Introduction and Progress

Phil Woodland

September 29th 2003

Cambridge University Engineering Department
Outline

• CUED EARS team

• CUED HTK Rich Audio Transcription Project
  1. Core STT Research
  2. Metadata Extraction
  3. Public HTK Development

• Progress to date

• Agenda
CUED EARS Team

Faculty Staff

- Prof Phil Woodland: Principal Investigator
- Dr Mark Gales
- Dr Thomas Hain

Research Associates

- Gunnar Evermann
- Dr Bin Jia
- Dr Do Yeong Kim
- Dr Marcus Tomalin
- Sue Tranter
- Dr Srinivasan Umesh
PhD Students

- Ricky Chan
- Andrew Liu
- David Mrva
- Khe Chai Sim
- Lan Wang
- Kai Yu
- Kit Thambiratnam (Junior Visitor)

Computer Officers

- Anna Langley
- Patrick Gosling [University funded]

Administrative

- Jane Hunter
Overall Aims

- 5 year project: CUED started July 2002

- Program aims for very large reductions in Speech-to-Text (STT) error rates for
  - Broadcast Data: initially Broadcast News (BN)
  - Conversational Telephone (CTS)
  - Multiple languages (English, Arabic, Mandarin)
  - Eventually real-time on standard hardware

- Augment STT to add extra information from audio (metadata)
  - information about speakers
  - aim for readable transcripts (by extracting slash units etc).

- Annual evaluations to evaluate progress
• WER targets for each year based on Progress Test Set (run once per year)
  – factor of 5 reduction in WER for CTS
  – factor of 3 reduction in WER for Broadcasts
  – move from unlimited computation to 1xRT
  – roughly 27% relative reduction in WER per year

• CUED HTK Rich Audio Transcription Project
  – Task1: Core Speech Recognition Technology
  – Task2: Metadata Extraction
  – Task3: Public HTK Development
  – Focus on English - also some work on Mandarin
Task 1: Core Speech Recognition

Improve/develop general techniques for speech recognition

Examine:

- Acoustic model training
  - discriminative training
  - adaptive training

- Acoustic model adaptation

- Language model adaptation

- Use of lightly supervised training techniques

- Optimal use of large amounts of training data

- System design and resource allocation
Task 1: CU-HTK Performance

Starting point for task 1 was the CU-HTK Hub5 system which had the lowest error rate on Hub5/Switchboard since 1998.

Progress on dev01 over the past few years:

<table>
<thead>
<tr>
<th>year</th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 full</td>
<td>19.3</td>
<td>32.5</td>
<td>33.2</td>
<td>28.3</td>
</tr>
<tr>
<td>2001 full</td>
<td>18.3</td>
<td>31.9</td>
<td>32.1</td>
<td>27.3</td>
</tr>
<tr>
<td>2002 full</td>
<td>16.4</td>
<td>29.2</td>
<td>27.4</td>
<td>24.2</td>
</tr>
<tr>
<td>2002 10xRT</td>
<td>18.3</td>
<td>31.9</td>
<td>31.0</td>
<td>27.0</td>
</tr>
<tr>
<td>2003 10xRT</td>
<td>15.8</td>
<td>26.9</td>
<td>25.9</td>
<td>22.8</td>
</tr>
</tbody>
</table>

- State-of-the-art WER reduced on Hub5 at avg of 8% relative per year

- For EARS must greatly accelerate this progress:
  - new algorithms
  - large amounts of low-cost training data.
Task 1: Discriminative Training

Recently developed discriminative training schemes are being widely adopted:

- Lattice-based framework for large vocabulary recognition
- Extended Baum-Welch optimisation
- Maximum Mutual Information Estimation - MMIE
- Frame discrimination
- Minimum Phone Error (MPE)
- Large reductions in WER
- Can make good use of large training corpora
Task 1: Discriminative training (2)

Improved generalisation via
- Acoustic deweighting - smoother posterior distribution
- I-Smoothing
- Appropriate complexity of language models

Work in EARS to consider:
- Alternative training criterion
  - Improved error-rate driven schemes
  - Further improvements for generalisation
- Interaction with lightly supervised data
- Learning other types of parameters:
  - Feature-space transforms
  - Model adaptation transforms
Task 1: Adaptive Training

- Focus on transcription of *found* data:
  - Non-homogeneous training (& test) data
  - Wide range of speakers/ channels/ acoustic noise
  - Pooling data into one block degrades separability

- Adaptive training
  - Use transformations to represent acoustic differences
  - Possible need for non-linear transformations
  - *Factor* the different types of variability

- Integrate discriminative and adaptive training
Task 1: Acoustic Factorisation

- Explicitly models all acoustic factors
- speaker/channel condition/acoustic environment
- From the training data extract
  - canonical model, given all the acoustic factors
  - distributions over factor transform parameters
- A generalisation of adaptive training
- Work will examine
  - forms of factor transformations (possibly non-linear)
  - forms for prior distributions for factor transformations
  - discriminative estimation schemes
**Task 1: Acoustic Model Adaptation**

Model adaptation will continue to be important. Focus on situations where no supervised adaptation provided: unsupervised transcription-mode.

- Adaptation provides improved performance for particular speakers / channel /acoustic environment, particularly when mismatched to training data.

Work will examine:

- Rapid adaptation using transform posteriors
- Extensions of lattice-based approaches for unsupervised adaptation
- Explicit models of acoustic factors
- Improved handling of unsupervised adaptation
**Task 1: Light Supervision**

- Need to use substantially more data than current systems (factor $>10x$ ?)
- Need for **low-cost** transcriptions e.g. closed captions/quick transcription
- Discriminative training yields larger improvements from extra data than maximum likelihood techniques.

Investigate schemes for handling lightly supervised training data

- Allow training schemes to account for transcription uncertainty
- Use of discriminative training schemes with low-cost transcriptions
- Interactions with adaptive training schemes
- LM training with low-cost transcriptions
Task 1: Language Model Adaptation

Need to incorporate domain specific knowledge especially in cases of limited LM training data

• Automatic determination of basis “themes” (e.g. topics/styles etc) for language model adaptation

• Basis themes define a continuous space

• Basis themes defined using e.g.
  – Latent Semantic Analysis
  – Probabilistic Latent Semantic Analysis
  – Automatic clustering schemes

• Adaptation by either
  – Interpolation of basis theme models
  – Linear transformations (?)
Task 1: System Design and Resource Allocation

Current full systems can be very complex and difficult to optimise. Typically

- large numbers of techniques combined
- multiple systems combined
- choices made ad-hoc/empirical

This work for EARS will

- predict performance of systems without full system evaluation
- determine optimal systems / resource allocation for particular portions of data
- investigate the generation of complementary systems
Task 2: Metadata Extraction

• Proposal divided metadata into
  – Acoustic: speaker and channel/noise conditions
  – Linguistic: (pseudo)-punctuation and topic

• Aim to extract metadata for two purposes
  – Feedback to help processes in speech-to-text (segmentation, clustering, etc)
  – To produced enhanced transcriptions

• To date worked on the tasks as they have been defined/refined
  – Diarisation: speech/non-speech detection and speaker tracking
  – Slash Unit (SU) detection/classification
Task 3: HTK Development – Background

• The HTK (Hidden Markov Model Toolkit) has been at the core of ASR research at Cambridge for more than ten years

• Very flexible, modular toolkit written in ANSI C

• CU-HTK: internal version used at CUED for research & system building

• HTK3: available for download since Sep 2000 http://htk.eng.cam.ac.uk

• More than 20000 users worldwide, heavily used in teaching, active support mailing lists

• Important features for state-of-the-art ASR systems missing in HTK 3.1

• Allows technology transfer to other groups

• Provide baselines for a range of tasks and testsets
Task 3: HTK Development

• Many of the advanced features present in internal CU-HTK are being extended, generalised and documented for inclusion into HTK3

• Other code are being implemented from scratch

• New features in HTK will include:
  – LM tools for training word/class-based n-grams
  – Lattice processing tools
  – Large vocabulary decoder for lattice generation and rescoring
  – Discriminative training tools
  – Improved adaptation routines, incl. adaptive training
Task 3: HTK-based infrastructure

- Make infrastructure available for standard tasks: WSJ, Switchboard, BN
  - task-specific “recipes” documenting standard way of building systems
  - baseline systems (e.g. MLE triphones), incl. HMMs, LM, dicts
  - word lattices for standard testsets, e.g. previous evals

- Provide support via mailing lists, both to novice users and experienced researchers

- Organise yearly meetings of HTK users to exchange experiences

- Include more advanced topics into the HTK-Book, e.g. class-based LMs, discriminative/adaptive training, large-scale system design
Progress to Date: CTS

• Much improved performance for CTS
  - Lowest error rates in the primary RT02 CTS unlimited computation
  - Lowest error rates in RT03 on the CTS Progress Test Set.

• New techniques (in RT02) included
  - heteroscedastic linear discriminant analysis (HLDA)
  - Minimum Phone Error training
  - discriminative adaptive training
  - use of single-pronunciation dictionaries (SPron)
  - 6-way system combination using triphone/quinphone versions of SAT-HLDA models, non-SAT Spron models and non-HLDA model
  - System lowest error rate in RT02 by a large margin
  - Many techniques taken up by other sites
• Improvements for CTS RT03 included
  – Automatic segmentation
  – Improved MPE training (lattice regeneration)
  – Variable number of Gaussians per state
  – Improved acoustic transcriptions
  – Increased training data
  – Improved language models
Progress to Date: Broadcast Speech

- Ported a range of techniques from CTS to BN (based on 1998 CUED BN 10xRT system)
  - HLDA
  - MPE
  - discriminative SAT
  - SPron
  - Confusion networks and CNC

- New discriminative MAP technique for gender-dependent models

- First 10xRT BN system to use system-internal system combination

- Lowest WER on BN English for RT03 Current and Progress Test sets (10xRT)

- Since RT03 developed lightly supervised discriminative training
Progress to Date: Fast Systems

- First HTK CTS 10xRT system developed for RT02
  - Good performance but many missing features from full-system

- Now developed unified architecture for such systems for BN and CTS
  - Includes many features from unlimited computation systems
  - “Fast-gap” much reduced
  - Much improved 10xRT CTS system: about 1% higher WER than 190xRT system

- For BN also submitted best 1xRT system in RT03 (first stages of 10xRT system)
Progress to Date: Mandarin

- Developed first Mandarin CTS system at CUED
- Similar to English but includes tone modelling and pitch features
- Fairly simple system developed for RT-03 evaluation
  - Small amount of acoustic training data
  - System had single “stream” with no system combination
- Since RT-03 developed improved system
  - Fixed error in use of training data
  - More complex systems investigated (including system combination)
  - Compared use of BBN dictionaries/phone sets/tokenisation
  - Current system now much closer to BBN’s RT-03 performance
Progress to Date: PhD Research

Ongoing work includes:

- Lightly supervised discriminative training
- Discriminative adaptive training
- Optimal resource allocation to design effective fast systems
- Advanced covariance and precision modelling which is so far yielding
- Acoustic factorisation which could yield very fast unsupervised adaptation
- Automatic complexity control and use of multiple projections
- Implicit topic-based adaptive language modelling

Will aim to include results from most promising work into future STT evaluation systems.
Progress to Date: Metadata

Diarisation

- Automatic CTS segmentation: 0.5% increase in WER relative to manual
- Improved segmentation for STT and greatly improved diarisation scores (halved since Jan 2003 dry-run)

SU Research

- Statistical language models + prosodic feature model (decision trees)
- 18% reduction in SU error compared to a simple baseline system developed for the Jan’03 dryrun
- System being prepared for RT-03f metadata evaluation
Progress to Date: HTK

- CUED internal HTK source code base used is now compatible with HTK3
- All future internal developments to HTK will be made available
- HTK 3.2 released including
  - New language modelling toolkit (HLM)
  - New set of lattice tools
  - Many bug fixes
- Complete rewrite of the HTK adaptation code which greatly increases flexibility and supported adaptation types
- Continued development for release next year of
  - discriminative training tools
  - large vocabulary decoder
Progress to Date: Infrastructure & Collaboration etc

- Established team of RAs and PhD students in Cambridge (in new offices!)

- Unified STT infrastructure for English CTS and (BN) transcription in terms of core codebase, data files, scripts etc.

- Served on EARS board and liaised with other EARS STT sites, LDC etc

- Collaboration with other sites including release of
  - CUED CTS transcriptions and lattices
  - shared creation of development test sets

- Inter-site technical meetings and discussions
  - EARS STT workshop (Martigny)
  - 2-day technical meeting with BBN/LIMSI
Progress to Date: Presentations, Reports & Publications

• Starting to have conference presentations on CUED EARS-funded work
  – 1 paper at ICASSP’03 Hong Kong
  – 4 papers accepted for ASRU’03 St. Thomas

• Substantial technical report on Metdata Diarisation work

• Many workshop presentations
  – RT02 and RT03 workshops
  – EARS PI meetings
  – Martigny STT workshop

• All of the above on CD-ROM provided (linked from EARS web site)
Agenda Monday 29th

9:00 Arrive & timetable for visit
9:15 Introduction to CUED EARS Project & Overview of Progress
10:00 Overview of CTS English systems from 2002 and 2003
11:00 tea/coffee break
11:15 Work on Broadcast News for Eval03 & Post-Eval Expts
11:50 Fast System Design
12:20 Large Scale Training Experiments with TDT Data
13:00 Lunch (LR4)
14:00 Work on Mandarin CTS Transcription
14:30 Metadata: Introduction
   Metadata: Diarisation
   Metadata: SU detection/Classification
16:00 tea/coffee break & EARS board meeting
16:40 SPron Modelling
17:15 HTK Development
17:40 Discussion of contract status/issues
Agenda Tuesday 30th

9:00   Talks by PhD Students
       Complexity Control
       Discriminative SAT
       Covariance Modelling
       Acoustic Factorisation
       Continuous Space Language Modelling
11:00  tea/coffee break
11:15  Tour of lab/facilities
11:30  Experiments with Quick Transcriptions & Fisher Data
11:45  General discussion
12:15  Future Plans & Wrapup
13:00  Lunch (LR6)

Aim of timings is to include enough time for questions/discussion