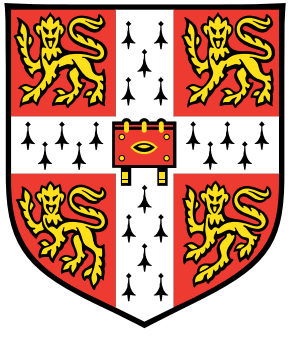
BBC Sample Data Processing

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Yanhua Long



University of Cambridge

April 24 2012

NST background

Natural Speech Technology

- Goal
  - to develop speech transcription models which can provide a wide domain coverage
    - from wide-ranging speech and text resources (lectures, meetings, TV drama, broadcast, etc)
- Initial work
  - get the first two sets of BBC sample data
    - BBC Radio4 and TV drama
    - BBC Reith lectures and Desert Island Discs (RL+DID)
  - initial investigations of the sample data
    - refine the manual transcriptions (time-stamp correction etc.)
    - build speech transcription systems

Preliminary data processing

Natural Speech Technology

- Train acoustic models and LMs on the two sets of sample data separately
- Automatic segmentation and speaker clustering
  - speaker labels are used to guide the acoustic model adaptation
- Two pass decoding
- Biased LM training
  - transcriptions are not accurate
  - interpolate the LM trained on BBC transcriptions with a general LM trained on other data resources (US-broadcast news), using interpolation weights heavily weighted towards the BBC LM.
  - to do the transcription refinement for improving segmentation quality and lightly supervised training.

Architecture of two pass decoding

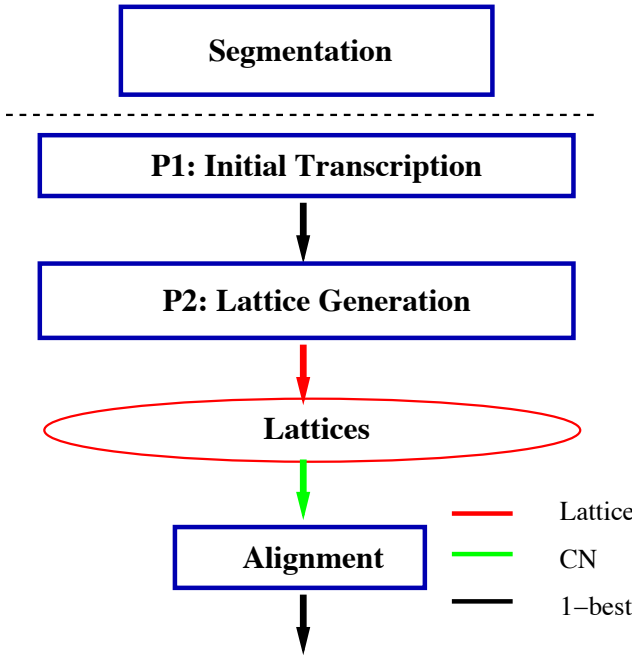


Figure 1: The architecture of two pass decoding

- Initially segment audio
- Initial hypotheses (P1) for adaptation of HMMs used in the P2-stage
- Unsupervised acoustic model adaptation
- Decoding and generate lattices (P2)

Processing BBC Radio4 and TV drama data

Natural Speech Technology

- Data sets
  - total: 25.56 hrs, 10% was taken as development data.
- Acoustic model training
  - US-BNE-rt04
    - CU HTK broadcast news acoustic model (bandwidth and gender dependent, discriminative training, adaptation, etc), trained on 1350 hrs data.
  - MPE-BBC
    - speaker adapted MPE acoustic model, trained on 16.59hrs data.
  - MPE-BBC-TSC
    - the same as MPE-BBC, but more data obtained from time-stamp correction, 20.52 hrs data.
- LM training
  - US-BNE: trained on a 59k word list, 1.4 billion word tokens.
  - BBC.train: trained on a 13k word list, 230k word tokens.
  - US-BNE+BBC.train: interpolated the above two LMs , 61K word list.

Processing BBC Radio4 and TV drama data

Natural Speech Technology

- Two pass decoding using multiple tri-gram LMs and acoustic models

Acoustic model	LM	WER% (show-level) dev-automatic		
		TV drama	radio4	all
US-BNE-rt04	US-BNE	73.51	28.50	44.45
MPE-BBC	US-BNE	69.40	23.08	39.51
MPE-BBC	US-BNE +BBC.train	**65.00**	**19.59**	**35.70**

- Segmentation using time-stamp correction (TSC)

Acoustic model	LM	dev-TSC		
		TV drama	radio4	all
MPE-BBC	US-BNE +BBC.train	56.96	18.30	32.02
MPE-BBC-TSC	US-BNE +BBC.train	56.46	17.48	31.31

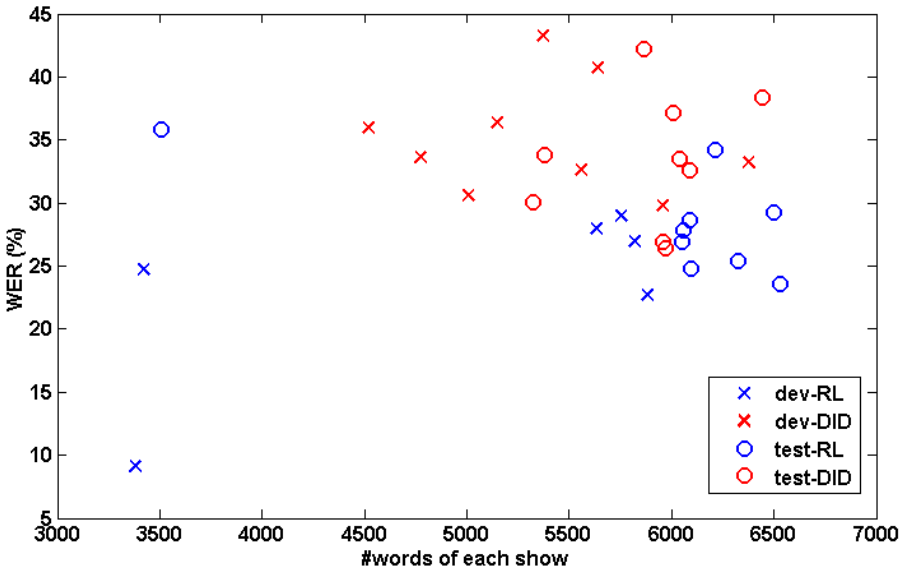
- Findings
  - An absolute WER reduction of 4.11% for TV drama and 5.42% for Radio4 gained from the MPE-BBC model, even it trained only on a small amount of data (16.59hrs compared with US-BNE-rt04, 1350hrs).
    - big acoustic environment mismatch between TV drama and broadcast news.
    - large accent mismatch between UK English (Radio4) and US English (broadcast news).
  - The quality of LM was also improved by introducing the small amount of vocabulary (2K) from BBC sample data, a 4.40% and 3.49% WER reduction were gained from the BBC.train LM.
  - After refining the segmentation
    - of development data, an absolute 8.0% WER reduction was achieved for TV drama, the development (test) data should be properly segmented.
    - of training data, more data was obtained to train acoustic models (+4 hrs), further performance improvements were obtained.

Work on Reith lectures and Desert Island Discs

- Data sets
  - Reith lectures (RL)
    - series of annual radio lectures, significant contemporary issues, delivered by leading figures
    - each show: one main speaker (the lecturer), presenter and few other people who join the discussion.
  - Desert Island Discs (DID)
    - a radio interview programme, one guest is invited in each week.
    - discussion about their imaginary stay on the island.
    - only two speakers in each show, the presenter and the guest, small portion of music.
  - total: *179.1* hrs, Desert Island Discs = *88.5* hrs, Reith lectures = *90.6* hrs
  - development + test sets: approximate 10% of the main speaker's data
- Properties of BBC manual transcriptions
  - no time-stamps for RL, approximate time-stamps for DID.
  - can not be directly used for acoustic model training.
  - have transcript insertions, deletions, and substitutions.

Initial evaluation on RL+DID

- Two-pass decoding for each show of the RL+DID dev and test sets
  - Using MPE acoustic model and interpolated LM trained from BBC radio4 and TV-drama
  - BBC manual transcriptions as reference, average WERs (%):
    - dev set: RL = 24.43, DID = 35.07; test set: RL = 28.10, DID = 33.52



- Reference transcript substitutions, insertions, deletions
- Only evaluate the decoding outputs in the audio regions which have manual reference transcript
  - still have reference transcript substitution and insertions

Work on BBC data at Edinburgh

Natural Speech Technology

- Can a system designed for meeting recognition help recognition of broadcast audio?
  - Investigated the use of the RT09 meeting recognition system developed during the AMI and AMIDA projects partly at Sheffield and Edinburgh.
- Findings: models trained on BBC data alone outperform the RT09 models...
  - However, using BBC training data to extract neural-network and fMPE features ("FM1") trained for RT09 system leads to much better performance compared to use of standard PLP features.
- Early results demonstrate that use of discriminative features can enable out-of-domain data to be used for rapid system development in a new domain.

Edinburgh's system results

Natural Speech Technology

- For models trained on BBC data, a simple one-pass decoding setup was used.
- Only ML-training.
- Initial CU segmentations were used for training.

Acoustic model	LM	WER% (show-level)		
		TV drama	radio4	all
AMI RT09	RT09	77.2	38.9	52.5
AMI RT09	RT09+BBC	76.6	33.3	48.6
BBC ML-PLP	RT09+BBC	78.7	28.8	46.5
BBC ML-FM1	RT09+BBC	**67.9**	**21.4**	**37.9**

Results are not directly comparable with CU results as a smaller (50K word) vocabulary was used for recognition.

Summary and future work

Natural Speech Technology

- Summary
  - Initial experimental investigations have been done on both of the two sets of BBC sample data which covers different domains.
    - different mismatch types among different domains.
    - large performance gap between TV drama and the other three types of data.
    - transcribing the data like TV drama tasks will be more challenging and interesting.
  - Initial results using Edinburgh's RT09 meeting recognition system demonstrated that:
    - the use of discriminative features can enable out-of-domain (meetings) data to be used for rapid system development in a new domain (BBC radio and TV drama).
- Future work
  - expect to get much more training data to have a wide domain coverage
    - wide range of programmes or more diverse data types
  - lightly supervised & unsupervised training from the approximate transcripts
  - domain-specific adaptive training & speaker adaptive training