The LIMSI CHIL RT'07 Lecture Transcription System

L. Lamel, E. Bilinski, G. Adda, J.L. Gauvain,

C. Barras, X. Zhu

RT07 meeting
Baltimore, MD
May 10-11, 2007

Overview

- Overview of the speech recognizer
- Acoustic model training
- Language model training
- Development and evaluation results
- Summary and some comments

Speech Transcription System Overview

- Same core technology as LIMSI Broadcast News transcription system
- Two components: Audio partitioner and Word recognizer
- Task-specific acoustic & language models and lexicon
- 35k and 58k word-lists
- Two-pass decoding
- Each decoding pass generates a word lattice which is expanded with a 4-gram LM
- Initial hypothesis generation with 3-gram LM, small cross-word position-dependent, gender-independent AMs
- MLLR adaptation (2 global regression classes)
- Runtime: about 20xRT

System Improvements

- Revised audio partitioner (X. Zhu talk tomorrow)
  - Iterative GMM clustering replaced by BIC clustering
  - GMM-based speaker identification clustering stage
  - Built GMMs on beamformed, tuning on RT06s
- Updated acoustic models
  - Lightly supervised training with TED speeches
  - ICSI beamforming software used to process lecture data (training and test)
  - MAP adaptation with beamformed data

Acoustic Front-end

- PLP-like analysis
- 8kHz bandwidth
- 39 features: 12 cepstrum coefficients and the log energy, 1st and 2nd derivatives
- Cepstral mean and variance normalization (by segment cluster)

Acoustic Training Data

2005 data: Almost 97h of IHM data from 4 sources

- **TED:** lapel mics, 39 speeches, 9.3h
- **ISL:** lapel mics, 18 meetings, 10.3h
- **ICSI:** head mounted mics, 75 meetings, 59.9h
- **NIST:** head mounted mics, 19 meetings, 17.2h

2006 data: additional 76h of data

- **ICSI:** FF tabletop, 75 meetings 70h
- **CHIL:** head mounted, 17 seminars, 6.2h

2007 data: additional 66h of data

- **TED:** lapel mics, 190 speeches, 46h (lightly supervised)
- **Beamformed data:** all available lecture data (trn+dev07+rt06s), $<$10h

TED

- Lapel mics, 190 speeches, 46h
- Lightly supervised training
- Data were transcribed with modified RT06 IHM system
- Biased LM trained on TED texts interpolated with RT06 LM
- WER between biased and RT06 LM: 8%
- Estimated WER: 5-20% (forgiving comparison)

Acoustic Models

- Context-dependent models, triphone with backoff
- Separate cross-word/word-internal statistics
- Tied states with decision tree, 152 questions (position, distinctive features, neighbors)
- 32 Gaussian mixtures per state
- Gender-independent models
- MLLT, SAT trained
- Small models in pass1: 5k contexts, 5.2k tied states, 165kG
- Larger models in pass2: 25k contexts, 11.5k tied states, 360kG
- AMs trained on pooled data, MAP adapted with beamformed data for mdm condition

Language Model Training Texts

- Audio transcriptions:
  - TED: 71k words
  - NIST: 156k words
  - ISL: 116k words
  - ICSI: 785k words
- Conversational telephone speech: 3M words
- Additional transcriptions: NIST RT04, RT05 data (57k words), AMI/IDIAP meetings (143k words), CHIL Jun04/Jan05 seminars (55k words), CHIL06 seminars: (38k words)
- BN transcriptions did not reduce perplexity (not used)
- Proceedings texts (20k articles, 46M words)

Lexicons

- 35k wordlist (RT05): optimized on jun04 and jan05 dev
- 20k words in audio transcripts, $\geq$3 times proceedings texts
- Case insensitive, 300 compounds, 1000 acronymns
- 34374 words, 43067 pronunciations
- OOV rate 0.61% on RT05s eval
- 58k wordlist (RT06): optimized on RT06 dev
- 75k most probable words selected by linear interpolation (8 seminar sources + proceedings, no CTS)
- Case sensitive, 2000 compounds & acronymns
- 57769 words,73480 pronunciations
- OOV 0.46% on dev data (RT05s eval)

Language Models

- Interpolation of 3 LMs trained on the text sources (weights 0.6, 0.3, 0.1)
  - Seminar and meeting transcriptions (1.42M words)
  - Proceedings texts (46M words)
  - Transcriptions of CTS data (29M words)
- 58k LM: 8.8M fourgrams, 19M trigrams, 5M bigrams
- 35k trigram: 6.6M fourgrams, 15M trigrams, 4M bigrams

Data set	*rt06*		*dev07*		*rt07*	
Language Model	OOV	Px	OOV	Px	OOV	Px
35k 4-gram	0.7	172	2.7	183	1.5	157
58k 4-gram	0.7	158	3.1	175	1.6	140
35k 4-gram renorm	0.4	157	0.9	165	0.7	136
58k 4-gram renorm	0.4	162	0.8	163	0.7	138

- Did not use Neural Network language model

Decoding Strategy

2 pass decoding

- Initial hypothesis generation with 35k trigram LM, small (5k contexts, 5k states) cross-word, position-dependent, gender-independent AMs (1xRT)
- Lattice rescoring with 35k 4-gram
- CMLLR and MLLR adaptation per segment cluster (2 global regression classes)
- Word lattice generation with 58k 2-gram LM and large (25k contexts, 11.5k states) cross-word position-dependent, gender-independent AMs
- Lattice expansion with 58k 4-gram LM
- Consensus decoding with pronunciation probabilities

Some Development Results

- Development data: dev07, rt06 eval (not representative)

dev07	*WER (%)*
Baseline, RT06 FF	64.4
Updated segmentation	64.0
Chunked data	63.0

rt06s bmf	*WER (%)*
rt06s mdm AM	65.2
+ pool bmf data	64.4
+ MAP with bmf data	62.2
+ tuning	61.0

Final AMs were MAP adapted using all beamformed data including rt06s

RT07s SASTT Results

scoring	*Cor (%)*	*Sub (%)*	*SpSub (%)*	*Del (%)*	*Ins (%)*	*WER (%)*	*SER (%*
SASTT	47.6	29.8	4.2	18.4	5.5	57.9	40.0
STT	51.8	29.7		18.4	5.6	53.7	38.3

Some Contrastive Results

scoring	*Cor (%)*	*Sub (%)*	*SpSub (%)*	*Del (%)*	*Ins (%)*	*WER (%)*	*SER (%*
SASTT	47.6	29.8	4.2	18.4	5.5	57.9	40.0
STT	51.8	29.7		18.4	5.6	53.7	38.3

- rt07s ihm results with SRI segmentations
  RT06s ihm AMs: 42.2%
  RT07s multistyle AMs: 40.6%
- rt07s sdm results
  60.7% STT
  63.9% SASTT

Summary

- System development mainly using RT06s lecture data and to some extent the dev07 data
- dev07 data is not representative of eval data
- Some development studies that did not pay off:
  - Reduced phone sets did not combine better than standard
  - Alternate pronunciations to model foreign accents
- Many model variants, tuning → all gave same result!
- No language model development this year (also did not use NN LM)
- Main improvement from use of ICSI beamforming and improved speaker diarization

Proceedings Texts

Total: 20k articles, 46M words

TED texts:	426 papers	929k words
ASRU'99-05:	427 papers	1140k words
DARPA'97-99,04:	119 papers	317k words
Eurospeech'97-05:	3485 papers	7650k words
ICASSP'95-05:	7831 papers	14318k words
ICME'00,03:	996 papers	2101k words
ICSLP'96-04:	3202 papers	7198k words
LREC'02,04:	891 papers	2553k words
ISCA+other workshops:	2333 papers	6077k words

Text Processing

- Convert ps and pdf files to text (similar to IRST)
- Remove undesirable data: email, addresses, mathematical formulas and symbols, figures, tables, references
- Remove special formatting characters and ill-formed lines
- Transformation of acronyms and compound words (AHD, name lists)
- Case sensitive (capitalization of first word)
  - Many words have multiple forms
  - Graph with all possible caseings, use special interpolated LM
- Keep information potentially useful for downstream processing