The LIMSI RT03 BN Systems

J.L. Gauvain, L. Lamel, G. Adda, L. Chen, H. Schwenk

RT03 meeting
Boston, MA
May 19, 2003

TALK OUTLINE

- LIMSI 2003 BN system overview
- Development set design
- BN English system
- BN Mandarin system
- Conclusions

BN SYSTEM OVERVIEW (English & Mandarin)

- Same partitioning as ’98 BN system
  - Iterative maximum likelihood segmentation/clustering procedure using GMMs and agglomerative clustering
- Updated acoustic and language models
  - 4 sets of tied state triphones (31k contexts, 11.5k states), 16 Gaussians per mixture
  - MMI training
  - 65k vocabulary, 4-gram LM
  - Use of TDT4 audio data with closed-captions for training
- Revised decoding strategy (same as dryrun03 system)
  - 2 step decoding

STT ENGLISH DEVELOPMENT SET

- No appropriate BN dev data available
- Selected 6 TDT shows from the second half of January 2001
  20010117_2000_2100_PRI_TWD
  20010120_1830_1900_ABC_WNT
  20010122_2100_2200_MSN_NBW (no captions available)
  20010125_1830_1900_NBC_NNW
  20010128_1400_1430_CNN_HDL
  20010131_2000_2100_VOA_ENG
- Selection criteria: representative WER and date
- Normalized closed-captions aligned with recognizer hypothesis
- Manual correction for scoring shared with BBN, CUED and SRI
- Verification marked commercials segments to ignore during scoring

ACOUSTIC MODELS

- PLP-like frontend, cepstral mean and variance normalization (by segment cluster)
- Triphone models (31k contexts, 16 Gaussian mixtures)
- Separate cross-word/word-internal statistics
- Tied states with decision tree
- Training data: ∼150 hours (1995, 1996, and 1997 Hub4 data) + ∼90 of selected TDT4 data
- Telephone and wideband models
- Gender-dependent models from SI seed models with MMI training

TRAINING TEXTS

- Old newspapers and newswires (1994-1999, 1.37G words)
- Recent newspapers and newswires (01/2000-31/01/2001, 54M words)
- BN data (1992-1998, 273M words)
- Manual transcripts of the HUB4 acoustic training data, old dev and eval sets (1.9M words)
- TDT2 and TDT3 captions and transcripts (1998, 9.6M words)
- TDT4 captions and transcripts (10/2000-15/01/2001, 2.2M words)
- CNN data from CNN archive (01/2000-15/01/2001, 12M words)
- **Wordlist:** selected using cutoffs for each source
  Minimize OOV on dev03 data
  Lexical coverage $\sim$ 99.5% on dev03

LANGUAGE MODELS

- 65233 words including compound words (300) and acronyms (1000)
- Language models: 2-gram, 3-gram and 4-gram
  - Development LMs trained all sources predating Jan 15, 2001
  - Interpolation coefficients minimize perplexity on Dev03
  - Eval LMs trained on all sources predating Feb 1, 2001
  - **RT03 LM:** 21M bigrams, 44M trigrams, 34M fourgrams

LM	PRI	ABC	MSN	NBC	CNN	VOA	Avg.
RT02	10.1	12.3	11.1	11.8	18.6	17.8	13.6
RT03	9.5	11.8	10.0	10.6	17.7	16.5	12.6

DECODING STRATEGY

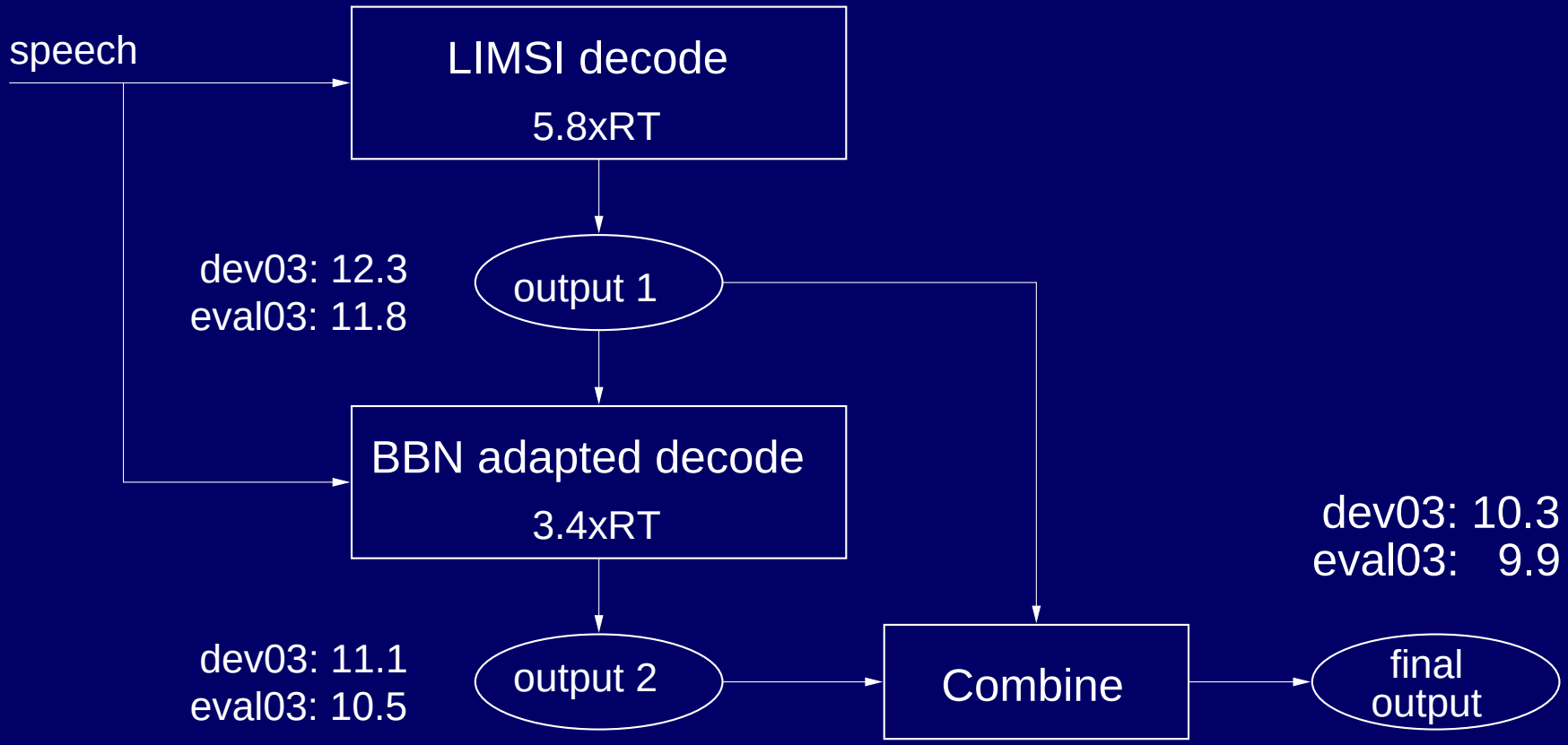
- Initial hypothesis generation with 3-gram LM, small cross-word position-dependent, gender-specific AMs (total 1.4xRT)
- Lattice rescoring with 4-gram
- MLLR adaptation and word lattice generation (2 global regression classes) with 2-gram LM and large cross-word position-dependent, gender-specific AMs
- Lattice expansion with 4-gram LM
- Consensus decoding with pronunciation probabilities

BN ENGLISH PROGRESS ON DEV03

RT02 system	14.5%
RT03 Dryrun system	14.1%
MMI training	13.6%
TDT4 LM	12.6%
TDT4 AM	12.2%
Optimized LM & decoding	11.8%

System	PRI	ABC	MSN	NBC	CNN	VOA	Avg.
RT02 (10x)	11.9	13.4	11.1	12.9	19.0	18.4	14.5
RT03 (10x)	8.6	11.0	9.6	10.0	16.7	14.8	11.8
BBN+LIMSI (17x)	8.2	9.0	8.3	8.9	14.7	12.5	10.3
BBN⊗LIMSI (9.2x)	8.0	9.2	7.9	9.0	14.9	12.9	10.3

BBN-LIMSI INTEGRATED SYSTEM (9.2xRT)



MANDARIN BROADCAST NEWS SYSTEM

L. Lamel, L. Chen, J.L. Gauvain

BN MANDARIN - OVERVIEW

- Same basic system as for English BN STT
- Modified audio partitioner for CBS/CTS (speech-in-noise GMM)
- Wideband & narrowband acoustic models
- Gender-specific, position-dependent triphones
- Lightly supervised acoustic model training
- 4-gram LM
- 57k wordlist includes all characters
- 2 pass decoding (1.4xRT + 8.4xRT)

ACOUSTIC MODEL TRAINING

- Hub4 Mandarin data from LDC (27 hours)
- 120 hours from TDT4 corpus
- Light acoustic model training: transcripts generated automatically with
  - AMs trained on LDC data
  - Source-specific LMs trained on TDT4 captions for Mainland sources (CNR, CTV and VOA) and the CBS Taiwan source
- CER about 7% on 4 CBS shows

ACOUSTIC MODELS

- Wideband models trained on Hub4-Mandarin and TDT4 Mainland sources (CNR, CTV, VOA)
- Narrowband models trained on narrowband version of above and TDT4 CBS data and 20 CBS shows (6 hours) with manual segmentations
- Gender-specific models
- Pass 1: 5500 contexts, 5500 tied-states, 16 Gaussians
- Pass 2: 21k contexts, 11500 tied-states, 16 Gaussians

LANGUAGE MODEL TRAINING

- Text sources available from LDC
  - TDT2,3,4 Mandarin transcripts (10.2M characters)
  - People Daily newspaper 1991-1996 (85M characters)
  - China Radio transcripts 1994-1996 (87M characters)
  - Xinhua news 1994-1996 (22M characters)
  - Acoustic training transcripts (0.43M characters)
- Text sources shared by BBN
  - People Daily newspaper 1997,1999,2000 (39M characters)
  - Central Daily News text 1997-2000 (61M characters)
  - CTS transcripts 1997-2000 (14M characters)

LEXICON

- 57707 words (including all characters)
- Essentially no OOVs
- 59152 phone transcriptions (2% alternate pronunciations)
- 61 phones including silence, fillers and breath
- 24 consonants
- 11 vowels, with 3 tones for each vowel (rising, flat and falling)

LANGUAGE MODELS

- Source specific language models (CBS, CNR, CTV, CTS VOA)
- Text segmentation using maximum match method
- Component LMs trained on each text source and each audio source
- Mixture weights chosen to minimize perplexity on Mandarin Dev03 data (shared by BBN)
- Weight of the audio transcript component set to 0.1.
- Minimum Discrimination Information adaptation for Taiwan sources (CBS, CTS) using the TDT4 CTS (0.66M chars) and CBS (0.46M chars) closed captions as adaptive data
- RT03 dev LMs trained on data through mid-Dec (predating Dev03 epoch)
- RT03 eval LMs trained on all data through Jan'03

LANGUAGE MODELS -CHARACTER PERPLEXITY

Show	*TDT LM*	*Source LMs*	*MDI-adapt*
CTV_MAN	191	167	-
CNR_MAN	248	204	-
VOA_MAN	274	249	-
CBS_MAN	508	412	390
CTS_MAN	623	495	460
Avg.	351	282	-

DEV'03 RESULTS

Show	Initial 3-pass SI	2-pass decoding			
		Common TDT4 LM	Source LMs + addl texts		
			SI	GD+wb/nb	TDT4 AMs
CTV_MAN	17.3	11.5∗	13.4	12.8	9.7
CNR_MAN	16.2	14.1	11.6	10.9	9.8
VOA_MAN	15.0	12.9	12.5	11.9	10.8
CBS_MAN	43.2	34.0	30.4	29.5	24.1
CTS_MAN	75.9	72.2	65.6	59.4	52.8
Avg.	34.5	30.2∗	28.0	25.8	22.6

∗ unfair LM for the CTS sources due to a naming reversal in captions

EVAL'03 RESULTS

Show	*Dev03*	*Eval03*
CTV_MAN	9.7	8.0
CNR_MAN	9.8	6.1
VOA_MAN	10.8	11.6
CBS_MAN	24.1	24.5
CTS_MAN	52.8	54.8
Avg.	22.6	21.7

CONCLUSIONS

- Updated BN systems for English and Mandarin
  - Improved acoustic models using additional TDT4 data
  - Improved language models (additional texts, improved smoothing)
  - WER reduction of 18% for English and 35% for Mandarin
  - CBS and CTS data are much more challenging than Mainland data accent? compression?
- Design of dev03 set for English
- Dev03 data are good indicators of eval performance