2003 CU-HTK Broadcast News English System Development

Do Yeong Kim, Gunnar Evermann, Thomas Hain,
David Mrva, Sue Tranter, Lan Wang, Phil Woodland,
and Rest of the HTK STT team

May 19th 2003

Cambridge University Engineering Department

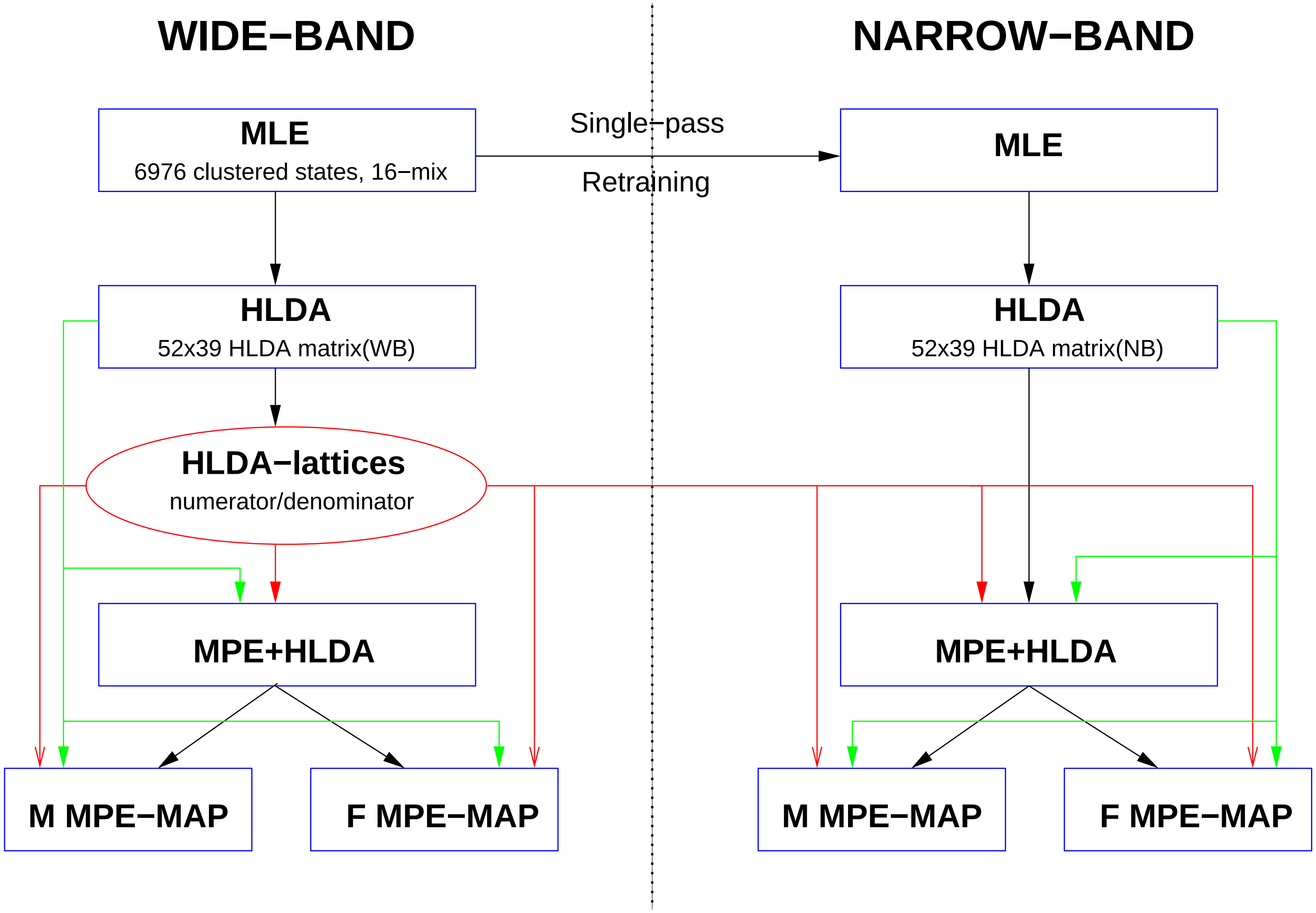
Overview

- Training data + Baseline Acoustic Models
- Adaptation Experiments
- Language Models
- Improved Acoustic Models
  - VarMix
  - Lattice-Regeneration MPE
- SAT
- SPron

Training data + Baseline Acoustic Models

- Training data : the 143 hours combined set of 1997 and 1998 data from LDC
  - **1997 data** 72 hours of acoustic BN training data
  - **1998 data** 71 hours of acoustic BN training data
- Front-end
  - 12 MF-PLP cepstral parameters + C0 and 1st/2nd derivatives + segment CMN (no VTLN or CVN)
  - Optional 3rd derivatives + HLDA
- Acoustic modelling
  - Decision tree state clustered, context dependent triphone models (6976 clustered states, 16-component mixture Gaussian)
  - Gender-dependent & bandwidth-dependent acoustic modelling
  - MLE/MPE/MPE-MAP training

Baseline Acoustic Models: Building Overview



Baseline Acoustic Models: Results (I)

- Development test sets

  BNeval98 two 1.5-hour data sets
  BNeval02 1-hour data set
  BNdev03 three hours of TDT-4 data from Jan '01 transcribed by STT sites

- 1998 CU-HTK BN-E LM (trigram)

- Single pass decoding without any adaptation

Baseline Acoustic Models: Results (II)

- HLDA transform
  - Estimate HLDA transform based on MLE baseline system
  - Add 3rd derivatives + HLDA, project 52 dim to 39
  - Consistent gain over different test sets, genders, and F-conditions

- MPE+HLDA
  - MPE training based on HLDA models
  - Significant gain over MPE or HLDA

	MLE	HLDA	MPE +HLDA
BNeval98			
F0	11.1	10.2	8.8
F1	20.1	18.5	15.5
F2	25.8	22.6	19.6
F3	20.9	19.1	17.3
F4	20.1	18.9	15.3
F5	28.1	27.2	19.1
FX	35.0	30.5	25.7
All	19.6	17.9	15.0
BNeval02			
All	17.9	16.0	13.6

%WER on BNeval98 & BNeval02

Basic Acoustic Models: MPE-MAP

- Gender-dependent discriminative training with MPE-MAP
  - Simple gender-dependent MPE model showed small gain (14.8%WER on BNeval98)
  - MAP-style update without losing advantage of discriminative training, see [Povey, Gales, Woodland: ICASSP2003]
- Most gains come from female speakers while both genders were improved

	MPE	MPE -MAP
BNeval98		
F	15.1	14.0
M	14.3	14.3
All	15.0	14.5
BNeval02		
F	14.8	14.5
M	13.3	12.5
All	13.6	13.0

%WER of MPE-MAP

Adaptation Experiments

	BNeval98			BNeval02		
	M	F	Total	M	F	Total
GI(HLDA+MPE)	14.3	15.1	15.0	12.9	15.3	13.6
1-best MLLR	13.8	14.4	14.4	12.0	14.1	12.6
Lat-MLLR 2trans	13.4	14.2	14.0	11.9	14.3	12.5
Lat-MLLR 2trans+FV	13.3	13.9	13.9	11.8	14.0	12.4
Lat-MLLR 4trans+FV	13.3	13.7	13.8	11.7	13.8	12.3

%WER for BNeval98 & BNeval02 after adaptation based on the GI unadapted models

- Apply global 1-best MLLR, phone-mark lattices, perform 4 iterations of Lattice MLLR

- By adapatation, WER was reduced by 8.7% relative on BNeval98, and 9.6% on BNeval02

- Small gains from FV and beyond 2 transforms

Improved Acoustic Model: Variable # of Gaussians

	BNeval98			BNeval02		
	F	M	Total	F	M	Total
HLDA	18.2	17.1	17.9	18.4	15.1	16.0
HLDA+VarMix	18.0	16.8	17.6	18.2	15.0	15.8

%WER on BNeval98 & BNeval02.

- Different number of Gaussians were assigned to each states according to the amount of available training data, while maintaining the average number of Gaussians per states the same as basic set-up (16 Gaussian/state)

- Marginal but consistent gains over two different test sets and both genders

Improved Acoustic Model: Lattice-Regeneration MPE

- Lattices for MPE training were regenerated using 4 iterations MPE+HLDA models with pruned bigram

- 4 more iterations of MPE with pruned bigram lattices and original lattices

	Total	F0	F1	F2	F3	F4	F5	FX	F	M
MPE+HLDA	15.0	8.8	15.5	19.6	17.3	15.3	19.1	25.7	15.1	14.3
Lattice-Regen	14.4	8.5	15.1	17.7	16.9	14.6	21.3	24.4	14.5	13.8

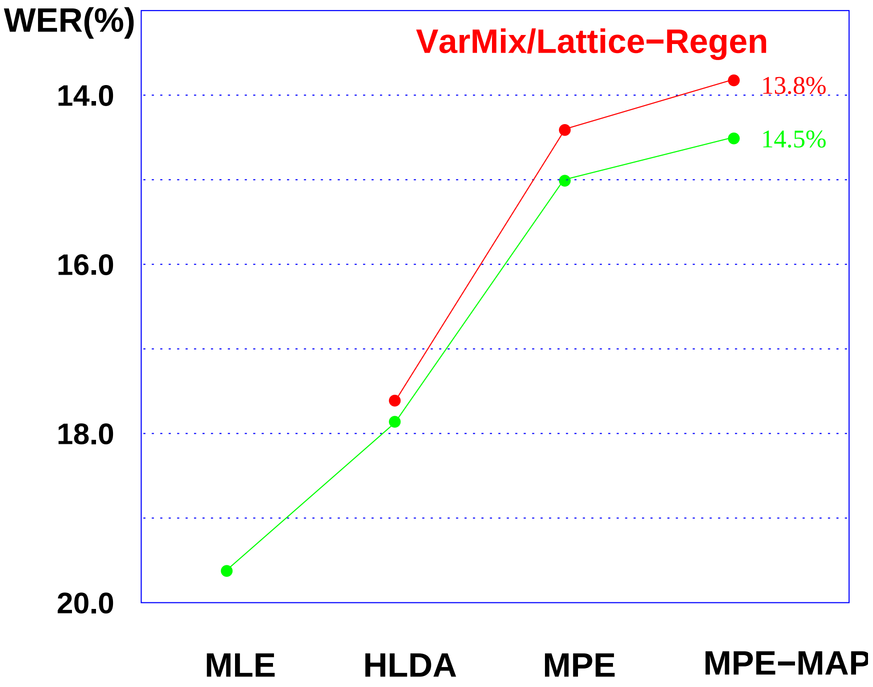
%WER of Lattice-Regeneration MPE on BNeval98

- Lattice-Regeneration MPE reduced 0.6% abs. error rates, and outperformed MPE+HLDA models in almost every F-conditions except F5 (speech from non-native speakers).

- Also works with gender dependent models (0.7% abs gain)

Improved Acoustic Model: Summary

- VarMix showed marginal gain
- VarMix/Lattice-Regeneration significantly recuded WER both in MPE(GI) and MPE-MAP(GD)
- 29.6% of relative reduction in WER (5.8% abs.) on BNeval98 by progress in acoustic modeling



%WER on BNeval98

Language Model (I)

- Language model training texts: 1,019 MW in total

- Subsets for interpolation

	Source	epoch	size (MW)
A	Primary Source Media BN transcriptions	1992-1999	275
	TDT 2 & 3 closed captions		
B	CNN shows transcription	1999-2001	66
C	TDT4 closed captions		2
D	broadcast news acoustic training transcriptions	1997-1998	2
	acoustic transcriptions for Marketplace shows	1996	
E	Los Angeles Times newswire service texts	1995-1998	674
	Washington Post newswire service texts	1995-1998	
	New York Times newswire texts	1997-2001	

No data from dates after mid January 2001 was used to conform with the epoch restriction for the eval data (Feb. 2001) and the BNdev03 set (late Jan. 2001)

Language Model (II)

- Wordlist
  - The 59k entry wordlist was chosen from BN LM training texts according to weighted sum of frequencies to minimize the OOV rate on BNdev03
  - 0.47% OOV rate on BNdev03
- Word-based language models
  - Good-Turing discounting with the HTK HLM toolkit on sets A, B, and E
  - Modified Kneser-Ney discounting with SRI toolkit on small sets C and D
  - All models merged into a single model
  - Entropy-based pruning
  - Pruned model has 8.8M bigrams, 12.7M trigrams, and 6.6M fourgrams

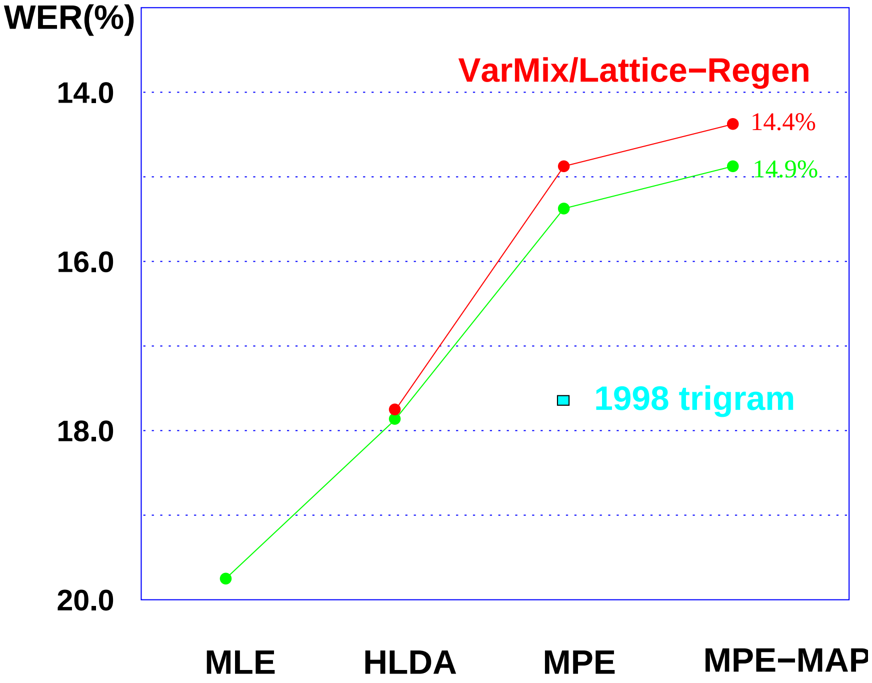
Language Model (III)

- Class-based trigram
  - Trained on broadcast material (sets A, B, C, and D) with HTK HLM
  - 1,000 automatically derived classes based on word bigram statistics
- Interpolation of word-based models with class-based trigram
  - The resulting word-based model was interpolated with the class-based model with weights of (0.87:0.13)
  - The interpolation weights were computed using EM
- Perplexities on BNdev03 with word-based trigram, fourgram, and interpolated fourgram with class-based trigram are 140.9, 121.5, and 119.1 respectively.

Improved Acoustic Model + New LM: Results on BNdev03

- Marginal gain by VarMix

- VarMix/Lattice-Regeneration approach showed consistent gain over previous MPE models



%WER of various acoustic models on BNdev03 with new LM(trigram)

SAT

	MPE-MAP+HLDA	SAT	SAT-VarMix
1-best MLLR	14.1	13.4	13.4
lat-MLLR 2trans	13.8	13.5	13.4
lat-MLLR 2trans+FV	13.6	13.3	13.0

%WER of SAT models on BNdev03

Note: All the experimental results here were obtained with an preliminary version of 2003 lanuage model(fg). Since we had WB SAT model only, NB results from MPE-MAP+HLDA 1-best MLLR was used to calculate %WER

- Show specific, gender-dependent clustering for test data
- SAT training used constrained MLLR
  - one transform for silence, another for speech
  - 5 iterations of interleaved transform and MLE model update
  - 6 iterations of MPE training with fixed transform

SPron

- Single Pronunciation dictionary
- Choose one pronunciation variant based on alignment of the training data
- Same approach as in CTS
- 6919 clustered states, 16 Gaussians/state, context dependent triphone gender-dependent / bandwidth-dependent acoustic modeling
- Acoustic model was built same way as MPron (MLE→HLDA→VarMix→Lattice-Regen-MPE→Lattice-Regen-MPE-MAP)
- Final GD SPron outperforms GD MPron by 0.5% abs. on BNdev03

Conclusions

- Successfully ported many techniques from CTS to BN
- Effective discriminative GD acoustic modeling using MPE-MAP
- Improved MPE performance by Lattice-Regeneration
- SAT: successful combination with MPE on BN
- SPron outpeforms MPron