Progress in Broadcast News English Transcription

D.Y. Kim, M.J.F. Gales, H.Y.Chan, P.C. Woodland,
S. Umesh & T. Hain

May 2004

Cambridge University Engineering Department

Overview

- VTLN & Linear VTLN
  - Unadapted single pass performance
  - Experiments in the "10xRT" system framework
- Acoustic model training using TDT4a
  - Unadapted single pass performance
  - CUED P1-P2 system results
- MPE training with a dynamic MMI prior
- New baseline development system (bnac+TDT4)
  - 10x framework performance
- RT03 `dev04` numbers (all others `dev03` (6 shows,3h) and `eval03` (6 shows,3h))

VTLN

- Speaker normalisation scheme
- Conventional VTLN:
  - warp frequency axis to normalise data
  - warp factors used estimated using ML:
    * commonly ignore effects of *Jacobian*
    * issue may be reduced using, for example, CVN
  - widely used in CTS, a few reports for BN task
  - awkward for BN as large number of segments
- Linear VTLN (LVTN):
  - approximate complex, non-linear warping by a linear transform
  - $Jacobian$ has a simple closed from solution
  - no need to re-parameterise the data
  - on-the-fly transformation (no issue for number of segments)
  - but only a linear transform (interaction with e.g. SAT)

Linear VTLN

1. $\boldsymbol{\lambda}^0$ is set to an appropriate non-VTLN model set $k = 0$.
2. Randomly select a subset of training data. For each warp factor α compute the set of warped feature vectors $\tilde{\mathbf{X}}^\alpha$.
3. For each α compute the linear transform (CMLLR), $\mathbf{W}^\alpha$, (block diagonal,, no bias, used for experiments)

$$\mathbf{W}^\alpha = \arg\max_{\mathbf{W}} \left(p(\tilde{\mathbf{X}}^\alpha; \boldsymbol{\lambda}^k, \mathbf{W}) \right)$$

4. For each segment of data estimate the warp factor

$$\alpha^{k+1} = \arg\max_{\alpha} \left(p(\mathbf{X}; \boldsymbol{\lambda}^k, \mathbf{W}^\alpha) \right)$$

5. Linearly warp the training data. A new model set, $\boldsymbol{\lambda}^{k+1}$, is then trained using single pass retraining and standard Baum-Welch estimation.
6. $k = k + 1$. Goto (3) until warp factors have stabilised.

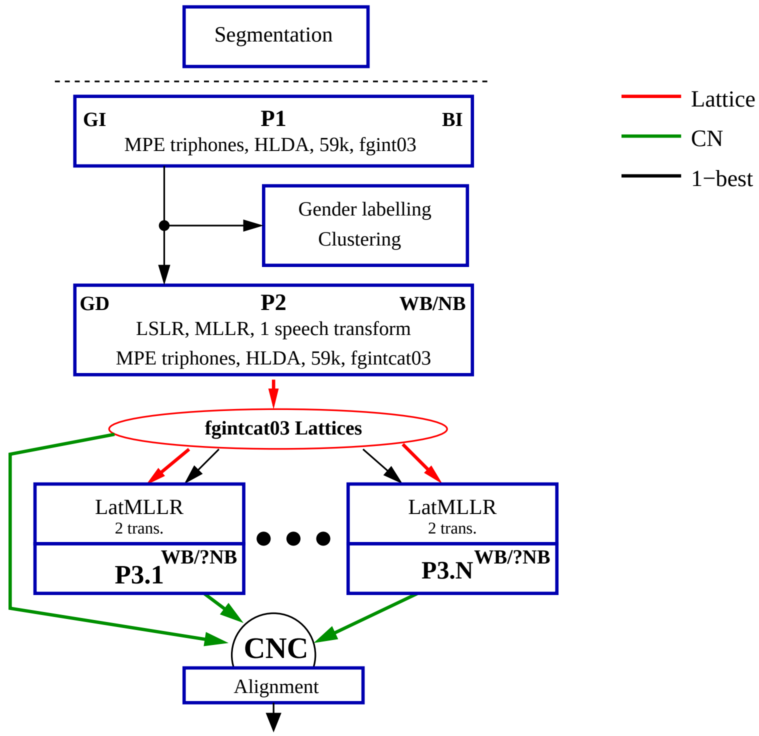
Unadapted (5x) Performance

Configuration		Front-end			
		CMN	CMVN	VTLN	LVTN
dev03	MLE	19.7	19.1	18.4	18.1
	+HLDA	17.9	17.9	16.9	17.1
	+MPE	15.2	15.3	14.6	14.6
	+GD	14.9	—	14.5	**14.4**
eval03	MLE	17.8	17.1	16.5	16.4
	+HLDA	15.9	15.9	14.9	14.9
	+MPE	13.7	13.7	13.2	13.0
	+GD	13.4	—	13.0	**12.7**

%WER with CMN (segment-based), CMN+CVN (CMVN), VTLN and LVTN cluster-based (> 500 frames) (no varmix, no lattice regeneration).

- Gain from CMVN disappears after HLDA
- VTLN & LVTN warp factors highly correlated (correlation coefficient of 0.98)
- VTLN & LVTN have consistent gains over CMN even for GD systems

10xRT framework



- Baseline RT03 system $N = 2$
  - 16 components/state for all systems
  - P3.1 SAT system generating WB only
  - P3.2 GD-SPRON system generating WB and NB
- Add GD-VTLN and GD-LVTN as P3.3/4 branches generating WB only

VTLN and LVTN in “10xRT” framework

	dev03	eval03
P3.0 (CMN)	12.3	11.2
P3.1 (SAT)	12.3	11.0
P3.2 (SPRON)	12.0	11.1
P3.3 (VTLN)	11.9	10.8
P3.4 (LVTN)	12.1	10.9
P2+P3.1+P3.2	11.6	10.6
P2+P3.2+P3.3	**11.3**	10.4
P2+P3.2+P3.4	11.4	10.5
P2+P3.1+P3.2+P3.3	11.4	10.4
P2+P3.1+P3.2+P3.4	11.4	**10.3**

%WER (varmix, lattice regeneration except SAT)

- Comparable performance for P3.3 VTLN and P3.4 LVTN
- Small but consistent gain after system combination

Acoustic Model Training using TDT4a

- TDT4a data set:
  - March-July 2001 (957 shows, ~ 530 hours raw data)
  - after advertisement removal and segmentation, 377 hours (71% of raw)

	bnac	TDT2	TDT4	TDT4a
ABC		47	25	49
CNN		183	42	201
NBC			23	52
MNB			34	75
PRI		71	52	
VOA		73	54	
Total	144	374	231	377

- Total 1,126 hours training data of bnac+TDT2+TDT4+TDT4a
- Currently no radio show transcriptions for TDT4a.

Unadapted (5x) Performance - MLE

- All models share the same decision tree and HLDA transform (bnac+TDT4)

	training data	hours	comp/state			
			16	20	24	28
dev03	bnac	144	17.9			
	bnac+TDT4	375	16.8			
	bnac+TDT4+TDT2	749	16.8	16.5	16.4	
	bnac+TDT4+TDT2+TDT4a	1126	16.7	16.3	**16.0**	16.1
	bnac+TDT4+TDT4a	752	16.7	16.3	16.1	**16.0**
eval03	bnac	144	15.9			
	bnac+TDT4	375	15.1			
	bnac+TDT4+TDT2	749	15.1	15.0	14.8	
	bnac+TDT4+TDT2+TDT4a	1126	15.0	14.7	14.5	**14.3**
	bnac+TDT4+TDT4a	752	14.9	14.7	14.4	**14.3**

%WER for eval03 (no varmix) single pass decoding with 03tg59k.

- Improved ML performance with additional data/components per state.

Unadapted (5x) Performance - 16 Component MPE

- All models share the same decision tree and HLDA transform (bnac+TDT4)

	hours	dev03	eval03
bnac	144	15.0	13.5
bnac+TDT4	375	13.8	12.5
bnac+TDT4+TDT2	749	**13.6**	**12.4**
bnac+TDT4+TDT2+TDT4a	1126	13.7	12.5
bnac+TDT4+TDT4a	752	13.9	12.5

%WER (no varmix, no lattice regeneration) for single pass decoding with 03tg59k.

.

- No gain by adding TDT4a data in unadapted 16 component configuration
  - performance on radio shows degraded

P1-P2 System 16 Component MPE Results

		P1	P2	P2-cn
dev03	bnac	15.9	12.9	12.6
	bnac+TDT4	14.5	12.0	11.8
	bnac+TDT4+TDT2	14.5	11.7	11.4
	bnac+TDT4+TDT2+TDT4a	14.3	11.6	**11.3**
	bnac+TDT4+TDT4a	14.2	11.4	**11.3**
eval03	bnac	14.9	11.9	11.5
	bnac+TDT4	13.6	11.1	10.9
	bnac+TDT4+TDT2	13.3	10.9	10.6
	bnac+TDT4+TDT2+TDT4a	13.0	10.5	**10.4**
	bnac+TDT4+TDT4a	13.0	10.7	10.5

%WER (no varmix, no lattice regeneration).

- Gains observed for configuration with adaptation (P2-stage)
  - 0.5%/0.4% gain from adding TDT4a to bnac+TDT4
  - little gain from using TDT2 data with TDT4a data

Dynamic MMI Prior

- Training data bnac+TDT4 (375 hours), 16 components per state.
- I-smoothing required for good generalisation of MPE:
  - standard scheme uses a *dynamic ML prior*
  - investigate IBM-style *dynamic MMI prior*
  - use *static GI-MPE prior* for GD models.

	dev03	eval03
MPE (dynamic ML prior)	13.9	12.6
+GD MPE-MAP	13.7	12.4
MPE (dynamic MMI prior)	13.6	12.5
+GD GI-MPE prior	**13.5**	**12.3**

%WER (varmix,no lattice regeneration) for unadapted (5x) single pass decoding with 03tg59k

- Consistent (small) gains with dynamic MMI prior;
- Consistent (small) gains with static MPE prior for GD modelling

2004 Baseline Development (Montreal) System

- Training data: bnac+TDT4 (375 hours)
- Acoustic model building differences to RT03 (gain)
  - no lattice regeneration and combination (approx -0.3%)
  - MPE training with dynamic MMI prior (approx +0.1%/0.2%)
  - GD trained with static GI-MPE prior (—)
- Same structure as RT03 10xRT system
  - All systems 16 components per state
  - P1: GI MPE model for initial transcription
  - P2: GD GI-MPE prior models for lattice generation
  - P3: SAT and SPRON models for lattice re-scoring
  - Confusion network decoding and system combination (P2+P3.1+P3.2)

Montreal 10xRT System Results

	dev03		eval03	
	RT03	Montreal	RT03	Montreal
P1	15.9	14.7	14.6	13.4
P2	12.7	11.9	11.6	10.8
P3.1 (SAT)	12.4	11.5	11.0	10.5
P3.2 (SPRON)	12.0	11.5	11.1	10.3
P2+P3.1+P3.2	11.6	**10.9**	10.6	**10.1**

%WER (varmix, no lattice regeneration).

- Comparison Montreal to RT03:
  - 0.8% absolute gain at P2 stage;
  - 0.7%/0.5% absolute gain at final system combination stage

Montreal + (a bit)

- Replace P1-P2 with best system:
  - use bnac+TDT4+TDT2+TDT4a (1126 hours, MPE ML-prior, non-varmix)
  - SAT and SPRON from Montreal system (375 hours MPE MMI-prior, varmix)

	`dev03`	`eval03`
P1	14.3	13.0
P2	11.6	10.5
P3.1 (SAT)	11.4	10.5
P3.2 (SPRON)	11.4	10.4
P2+P3.1+P3.2	**10.7**	**9.9**

%WER (no lattice regeneration).

- Comparison to Montreal system;
  - 0.3% better at P2 stage
  - 0.2% final improvement on both `dev03` and `eval03`
  - under 10% for `eval03` ...

RT03 dev04 Results (Reference)

- RT03 system was run on new dev data candidates.

show	Corr	Sub	Del	Ins	Err
20010125+2000+2100+PRI+TWD	89.6	8.1	2.3	2.2	12.6
20010127+1830+1900+ABC+WNT	88.5	8.9	2.7	2.0	13.6
20010128+1400+1430+CNN+HDL	84.4	9.6	5.9	1.8	17.3
20010130+1830+1900+NBC+NNW	88.2	8.3	3.4	1.4	13.1
20010130+2100+2200+MSN+NBW	91.2	6.2	2.7	1.0	9.9
20010131+2000+2100+VOA+ENG	88.9	9.0	2.1	2.7	13.9
Total	88.7	8.2	3.1	1.8	13.2

%WER for dev04 candidate shows.

- GLM and scoring script for RT03 were used for scoring.

Conclusion

- VTLN yields small gain for BN-E
  - LVTN comparable performance with conventional VTLN
  - simpler implementation for large BN systems
- A gain of 0.1%-0.2% using MPE training with a dynamic MMI prior.
- New Baseline development system (Montreal) bnac+TDT4:
  - gave 0.7% on `dev03` and 0.5% on `eval03` over RT03
  - combined with "best" P1/P2 system (1126 hours) additional 0.2%
- Lots of things to do:
  - investigate additional components per state/training data
  - incorporate new approaches e.g. precision matrices, DLT, CAT/ST etc
  - update language model using additional data