2003 CU-HTK English CTS Systems

Phil Woodland, Ricky Chan, Gunnar Evermann, Mark Gales,
Thomas Hain, Do Yeong Kim, Andrew Liu, David Mrva,
Dan Povey, Sue Tranter, Lan Wang, Kai Yu

May 19th 2003

Cambridge University Engineering Department

English CTS Development

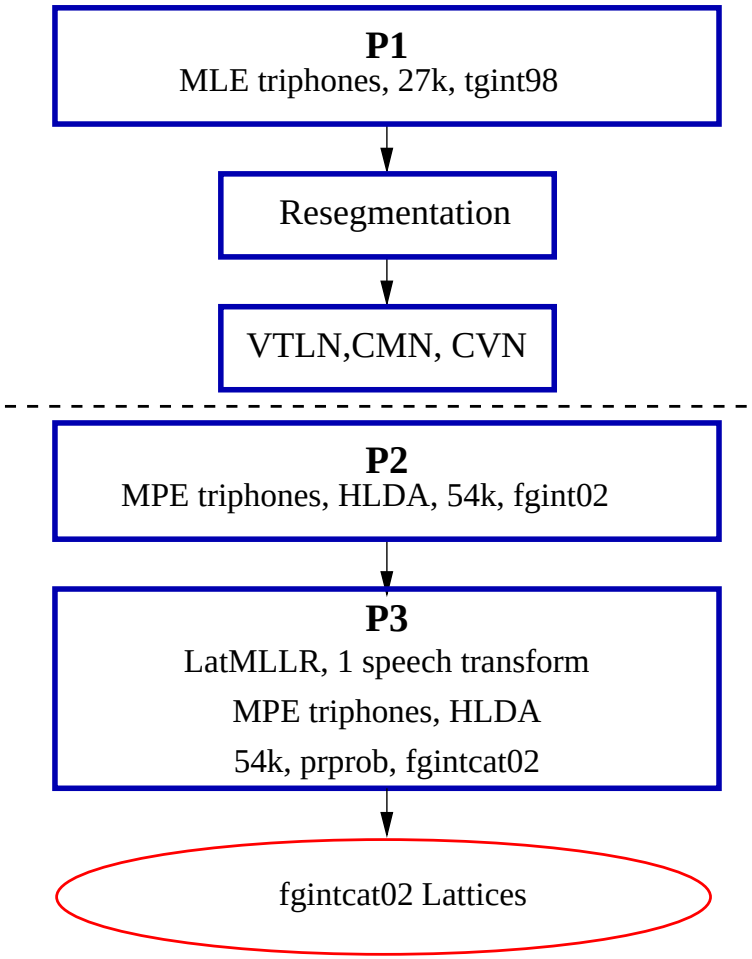
- 2002 unlimited computation system
- Training and test data sets
- New/Revised components
  - automatic segmentation
  - revised transcriptions
  - variable number of Gaussians
  - lattice generation for MPE training
  - SAT experiments
  - additional acoustic training data
  - SPron experiments
  - revised language models
- 2003 system performance
- Conclusions

2002 System

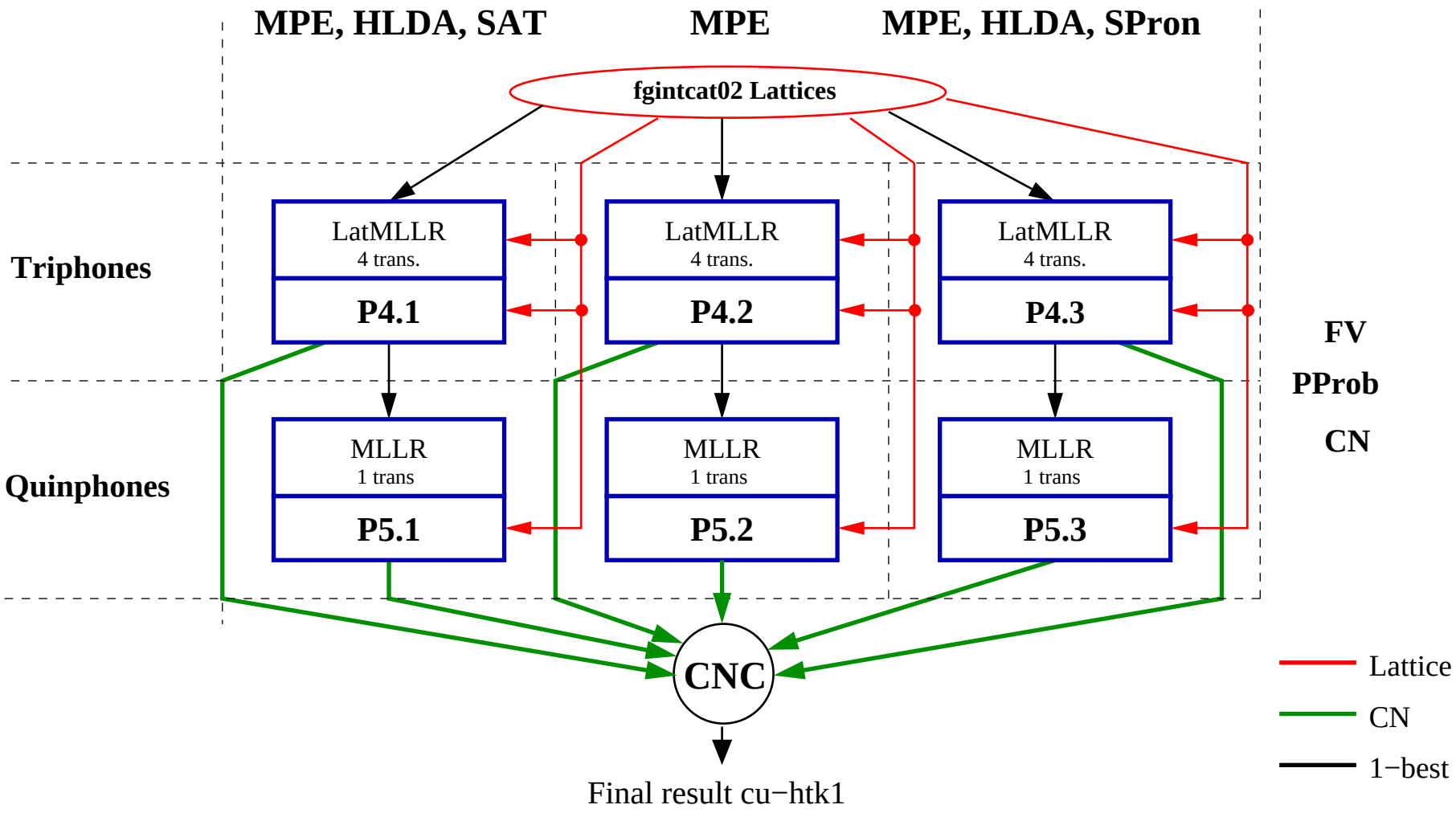
- Assumes manual segmentation into turns
- PLP, side-based CMN/CVN + 1st/2nd Δs (+ 3rd Δs & HLDA to 39 dims)
- Initial passes generate transcriptions for VTLN & initial adaptation
- Generates lattices with adapted triphone models and a bigram LM
- Expands the lattices to 4-gram plus trigram category model
- Rescores the lattices with adapted triphone and quinphone models
  - MPron HLDA SAT MPE triphone/quinphones
  - SPron HLDA non-SAT MPE triphones/quinphones
  - MPron non-HLDA non-SAT MPE triphones/quinphones
- Use confusion networks to represent each rescoring pass output & confusion network combination for highest posterior prob words and confidence scores

2002 System - Lattice Generation

- MLE P1 models
- MPE triphone models for P2/P3
- 28 mixture components (28 mix)
- HLDA
- Adaptation for P3 via Lattice MLLR
- Pronunciation probabilities
- HTK decoder HDecode



2002 system – Rescoring & Combination



Results on eval02 set

		Swbd1	Swbd2	Cellular	Total
P1	trans for VTLN	35.6	44.6	50.5	44.0
P2	trans for MLLR	24.6	30.9	34.8	30.4
P3	lat gen	22.5	28.0	31.3	27.5
P4.1	SAT tri	21.6	26.3	29.6	26.1
P4.2	non-HLDA tri	22.3	27.4	31.2	27.2
P4.3	SPron tri	21.5	26.6	29.1	26.0
P5.1	SAT quin	21.5	25.5	28.6	25.4
P5.2	non-HLDA quin	22.4	26.7	30.7	26.9
P5.3	SPron quin	21.5	26.4	28.8	25.8
CNC	P4.[123]+P5.[123]	19.8	24.3	27.0	23.9

%WER on eval02 for all stages of 2002 system, manual segmentation

- final confidence scores have NCE 0.289

Training and Test Data Sets

h5train02 248 hrs Switchboard (Swbd1), 17 hrs CallHome English (CHE) + LDC cell1 corpus (without dev01/eval01 sides) extra 17 hrs of data

h5train03 290 hr set. As above plus extra 12 hours of Switchboard I from final MSU transcripts

h5train03b 360 hr set. As above plus extra Switchboard Celluar I and Swd2 Phase2 data as released by BBN (CTRANS transcribed)

Development test sets

dev01 40 sides Swbd2 (eval98), 40 sides Swbd1 (eval00), 38 sides Swbd2 cellular (for manual segments)

eval02 40 sides of Swbd2; 40 sides of Swbd1; 40 sides of Swbd cellular. Can be used with manual or automatic segments

Automatic Segmentation

- Need to automatically segment the input data this year
- Used models with Gaussian mixture modes specific for cellular/non-cellular & male/female (256 Gaussians for male/female; 128 for silence)
- Constrained to have only one type of speech per side
- More details in diarisation talk

	Diarisation score (dryrun data)	% WER (eval02)
CUED dryrun segments	13.09	27.8
CUED sys03 segments	8.55	27.3
STM segments	39.89 (!)	26.7

Recogniser used in 10xRT system from Dec'02 (dryrun)

Revised Transcriptions

A mistake in the Switchboard training transcriptions used in building all CUHTK CTS systems since 2000 was discovered.

- Error in processing MSU Swbd training transcripts
- Some fairly common words systematically deleted (3% of tokens)
- Affected both acoustic models and LMs
- Rebuilt transcriptions based on final version of MSU transcripts
- Added 294 new conversation sides
- Rebuilding acoustic models only, for 2002 10xRT system on eval02 (manual segs), reduced WER by (only?) 0.5% abs (27.2 to 26.7)

Var #Gauss per state

- CU's std approach was N Gaussians per speech state and $2N$ for silence
- Set #Gauss as a function of number of frames γ_j available to train state j
- Use #Gauss $= k\gamma_j^p$, where p is a small power (e.g. 1/5)
- k is a normalising constant set to make the average #Gauss equal to N
- On CTS typically gives a 0.1-0.4% abs reduction in WER (see later tables)

SAT/Adaptation Experiments

- SAT tries to remove inter-speaker variability in training set by means of linear transform
- Use constrained MLLR to generate a single transform per training side (can operate in feature space)
- Interleave update of adaptation matrices and MLE HMM updates
- Perform MPE training based on SAT models with fixed transforms
- 0.3% abs improvement from SAT

	SAT				non-SAT			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
1 best std MLLR	17.7	31.1	30.5	26.4	17.7	31.6	31.0	26.7
lattice MLLR/FV	17.4	30.4	29.7	25.8	17.5	30.9	30.2	26.1

% WER dev01 manual seg 2002 fgintcat LM, HLDA MPE-trained triphones

Lattice-Based MPE Training

- Minimum Phone Error training (Povey & Woodland, 2002)
- Uses lattice-based training developed for MMI and extended B-W updates
- Includes "I-smoothing" of discriminative statistics with ML counts
- Requires the generation of lattices for the training set:
  - Correct transcription (corresponds to MMI numerator)
  - Representation of the confusable model sequences (MMI denominator)
- Denominator lattices generated in two steps
  - Word level lattice generation (uses training-data bigram LM)
  - Model-marking of HMM sequence and segmentation points (unigram LM)
  - Training procedure treats segmentation points as truth
  - Lattices generated using ML models (non-HLDA)

Modified Lattice-Based Training

- In 2002 no re-alignment/regeneration of lattices during discriminative training
  - In 2001 re-generated model-marked lattices part way through MMI training
- Now use heavily pruned training data bigram for word lattice generation
  - larger "denominator" lattices
  - better representation of confusable data
  - use pruned bigram scores in MPE training also
- Use HLDA ML models to generate lattices (rather than non-HLDA lattices)
- After 4 iterations of MPE training regenerate word and model-marked lattices with MPE models and use $both$ of lattices (combining at statistics level).

MPE Training with Modified Lattices: Results

	Swbd1	Swbd2	Cellular	Total
non-HLDA lattices	20.5	35.3	34.7	30.1
HLDA full bg + ug	20.4	34.7	34.3	29.7
HLDA pruned bg	20.0	34.4	34.0	29.4
MPElattice regen/comb	19.4	34.0	33.6	28.9

% WER dev01 manual seg 2002 trigram LM, unadapted 28mix HLDA triphones, 290hr training, MPron

- HLDA ML models to generate lattices reduces WER by about 0.4% abs
- Larger lattices with pruned bigram reduce WERs by about 0.3% abs
- This lattice regen/comb gives a further 0.5% abs improvement in WER

Additional Acoustic Training

- New Swbd2 data transcriptions provided by BBN (70 hours)
- About 1% abs reduction in WER for MLE HMMs and 1.3% for MPE
- Largest improvement for cellular data (2.2% abs) and Swbd2 data (1.4% abs)

	290hr train				360hr train			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
16 comp MLE	24.9	39.8	39.6	34.7	24.7	39.4	38.5	34.1
28 comp MLE	24.0	39.0	38.1	33.6	23.6	38.1	36.8	32.7
Var comp (28) MLE	23.9	38.8	38.0	33.5	23.1	37.8	36.8	32.5
MPE (8its)	20.0	34.4	34.0	29.4	19.4	33.2	31.7	28.0
MPE lat combine	19.4	34.0	33.6	28.9	19.0	32.6	31.4	27.6

% WER dev01 manual seg 2002 trigram LM, unadapted HLDA triphones

SPron Dictionary

- Modified procedure from 2002 CUHTK CTS eval system (Hain, 2002)
- Systematically remove all pronunciation variants
- If words were observed in the training data
  - Selection is based on pronunciation variant frequency
  - DP alignment of pronunciation variant pairs followed by merging variants with substitutions only and then phoneme deletions/insertions
- Training of statistical model on decisions above
  - For a pair of pronunciation variants identify target and source
  - Model uses phoneme substitution probs
- Unobserved words
  - Identify source variant from statistical model
  - Select primary variant by pairwise exclusion

SPron experiments

- Rebuilt SPron models with MPE lattice comb from MPron word lattices
- Lattice combination helps 0.8% with SPron models built like this
- Final MPron and SPron WERs very similar (SPron 1% abs better for MLE)

	MPron				SPron			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
16 comp MLE	24.7	39.4	38.5	34.1	24.4	38.3	37.5	33.3
28 comp MLE	23.6	38.1	36.8	32.7	23.0	36.9	36.1	31.9
Var comp (28) MLE	23.1	37.8	36.8	32.5	22.6	36.6	35.6	31.5
MPE (8its)	19.4	33.2	31.7	28.0	19.9	33.1	32.2	28.3
MPE lat combine	19.0	32.6	31.4	27.6	19.0	32.2	31.6	27.5

% WER dev01 manual seg 2002 trigram LM, 360hr training, unadapted HLDA triphones

- After eval03 found SPron lattices from scratch (new word lattices/model marked lattice + MPE regen) helps by only another 0.1% absolute

2003 language models

- Training data in 5 portions:
  - Revised MSU transcripts + CHE [3MW]
  - broadcast news setup (BN transcripts from PSM; CNN data; TDT data) [427MW]
  - Cell1 transcriptions [0.2MW]
  - Swb2 transcriptions from BBN/CTRANS [0.9MW]
  - google data from U of Washington [62MW]
- Used dev01, eval01 and eval02 as dev set
- Selected 30k words from acoustic transcripts plus top 54k words from BN (58k total). OOV rate 0.19% on dev set
- Trained 5 component 4-gram LMs; one class 3-gram LM

- "Small" text sources trained using modified Kneser-Ney (SRI LM); large text source using Good-Turing (HTK HLM)

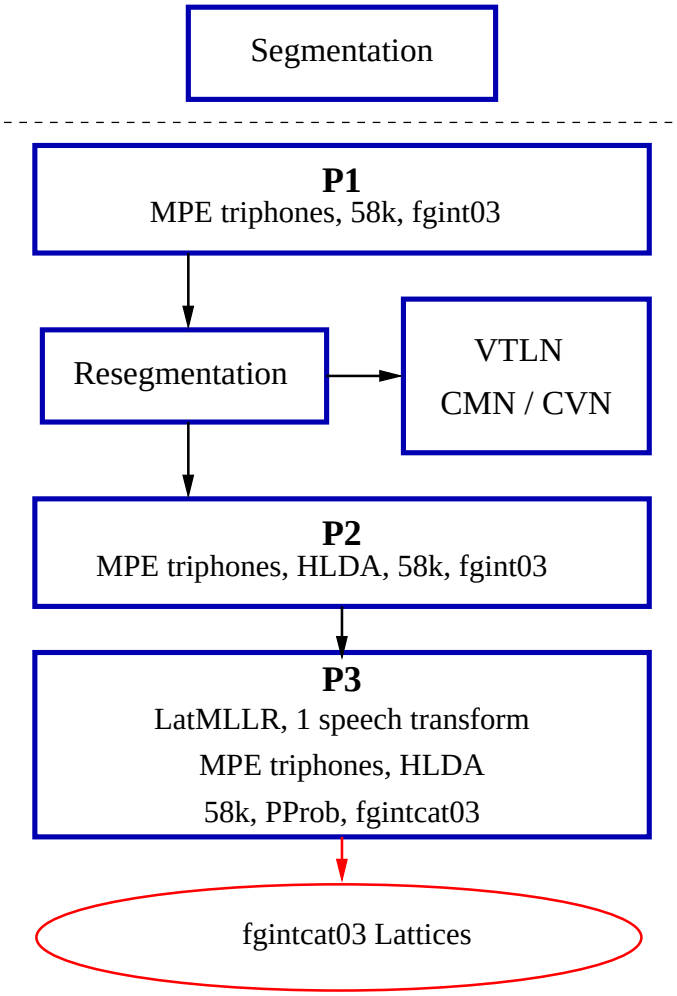
- 2003 merged fgintcat has 4.3% rel reduction in PP over 2002 model. With cat models the difference is 3.5%

- Effect of component 4-gram word LMs

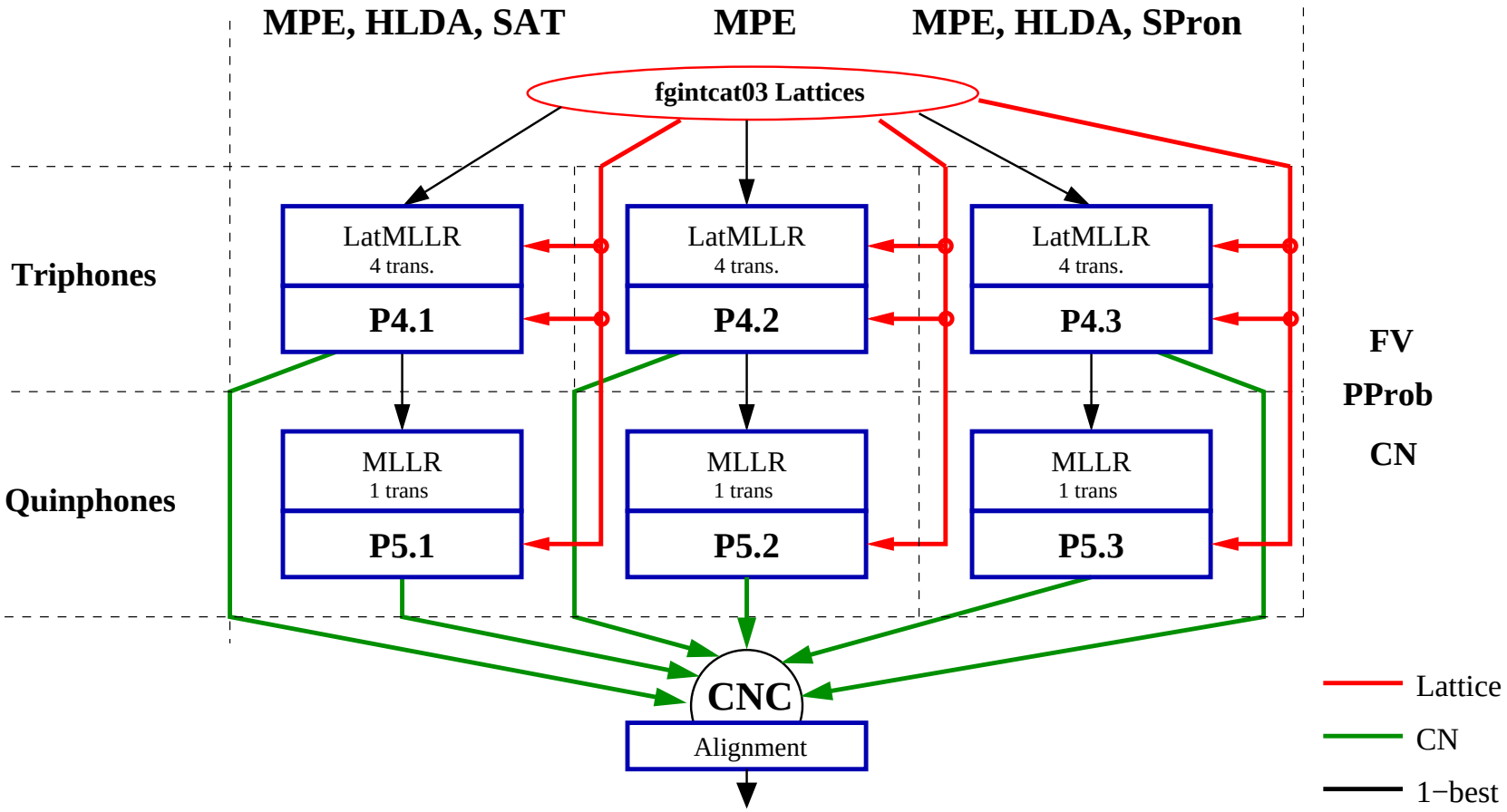
component LMs	fg PP
all	65.2
all minus google	65.9
all minus cell1	67.4
all minus swbdII	68.4
all minus che+swbdI	68.6
all minus BN+TDT+CNN	68.9

2003 System

- Automatic Segmentation
- Revised non-VTLN HTLDA MPE P1 models (290hr) + fg LM
- Revised MPE training for all other models (360hr)
- Modified SPron models for tri/quin
- Pronunciation probabilities in tri/quin
- Adaptation & system combination same
- Final alignment step



2003 System Part II



2003 System Performance (Eval02)

		Swbd1	Swbd2P3	Cellular	Total
P1	trans for VTLN	27.2	34.8	39.5	34.2
P2	trans for MLLR	23.6	28.9	31.7	28.4
P3	lat gen	21.1	25.1	27.6	24.8
P4.1	SAT tri	19.9	23.3	25.2	23.0
P4.2	non-HLDA tri	21.2	24.9	27.7	24.8
P4.3	SPron tri	20.4	23.7	25.6	23.4
P5.1	SAT quin	20.0	23.6	25.0	23.0
P5.2	non-HLDA quin	21.2	24.9	27.1	24.6
P5.3	SPron quin	20.1	23.9	25.3	23.3
CNC	P4.[123]+P5.[123]	18.6	22.3	23.7	21.7

%WER on eval02 for all stages of 2003 system (auto-segments)

Final NCE is 0.304

2003 System Performance (Eval03)

		Swbd2P5	Fisher	Total
P1	trans for VTLN	37.7	27.9	33.0
P2	trans for MLLR	31.8	22.6	27.4
P3	lat gen	27.5	19.3	23.5
P4.1	SAT tri	25.4	18.2	21.9
P4.2	non-HLDA tri	27.4	19.6	23.7
P4.3	SPron tri	25.6	18.5	22.2
P5.1	SAT quin	25.5	18.4	22.1
P5.2	non-HLDA quin	27.5	19.6	23.7
P5.3	SPron quin	25.7	18.7	22.3
CNC	P4.[123]+P5.[123]	24.1	17.1	20.7

%WER on eval03 (current test) for all stages of 2003 system (auto-segments)

Final NCE is 0.318

Conclusions

A number of changes and improvements have been made to the system although basic structure the same as 2002 system

- Automatic segmentation now gives only 0.6% increase in WER
- On eval02 data got 23.9% WER in 2002 with manual segments: now 21.7% with automatic segments. Approx 12% reduction in WER if use consistent manual segments
- revised Swb1 transcriptions: 0.5% abs
- variable number of Gaussians per state: 0.3% abs
- new MPE lattice generation/regeneration procedure: 1.2% abs
- new Swb2 data: 1.3% abs unadapted / no system combination