CU-HTK RT03 Mandarin CTS System

Bin Jia, Khe Chai Sim, Mark Gales, Thomas Hain,
Andrew Liu, Phil Woodland, Kai Yu & the HTK STT Team

May 19th 2003

Cambridge University Engineering Department

Mandarin CTS 2003 System

- Acoustic and Language Model Training Data
- Mandarin Phone Sets
- Tonal Decision Tree Questions
- Vocal Tract Length Normalisation and Pitch
- Varmix and MPE training
- Results

Acoustic Training Set-Up

- Acoustic/Training Test Data:
  - training data: 34.9 hours, 379 sides, from LDC CallHome (22.4hrs) and CallFriend (12.5hrs), 451K Words (+7K English word), 628K Characters
  - development data: `dev02` 1.94 hours from CallFriend
- Front-end
  - Reduced bandwidth 125–3800 Hz
  - 12 PLP cepstral parameters + C0 and 1st/2nd derivatives
  - Side-based cepstral mean and variance normalisation
  - Optional vocal tract length normalisation in training and test
  - Optional pitch (and derivatives) obtained from ESPS
- Acoustic Models
  - Gender independent models
  - Decision tree state clustered, context dependent triphones
  - Approximately 3000 distinct states

Language Model

- Sources of data (using LDC character-to-word segmentor)
  - Acoustic training data (modifier Kneser-Ney)
  - News corpora: TDT[2,3,4], China Radio, People's Daily, Xinhua (Good-Turing)

- Word LMs - 11K vocabulary, 0.17% OOV on dev02

Data	Bigram	Trigram
Acoustic	206.6	190.8
Acoustic+News Corpora	199.6	179.8

Perplexity results on dev02

- Class-based LM - 75 classes trained on acoustic trnascriptions

LM	Bigram	Trigram
Class	196.1	190.1
Class+Word	188.3	172.1

Perplexity results on dev02

Mandarin Phone Sets

# Phone Set	CER (%)
59-phone	58.1
46-phone	57.0

%CER for dev02 using 12 mix comp VTLN MLE trained systems and word trigram LM

- Two phone sets considered:
  - 59-phone set, start with LDC 60 phone set, remove tone markers and

    u:e → ue
  - 46-phone set, start with 59-phone set and split long final phones, e.g.

    [aeiu]n → [aeiu] n

    [aeio]ng → [aeio] ng

    uang → ua ng
- Mapping reduced CER by 1.1% absolute
- 46 phone set was used for all further experiments

Tonal Decision Tree Questions

Tonal Questions	CER (%)
×	57.0
√	55.7

%CER for dev02 using 12 mix comp VTLN MLE trained systems and word trigram LM

- Tonal questions incorporated into decision tree process (without pitch features):
  - 3% of possible questions were tonal
  - all tonal questions used for at least one tree
  - tonal questions normally used near top of decision tree
- Yields about 1.3% absolute reduction in character error rate
- Tonal questions were used for all further experiments

VTLN/Pitch Results

VTLN	Pitch	12 Comp	+HLDA	+Pitch
×	×	57.5	56.1	—
×	√	57.0	56.2	—
√	×	55.7	53.8	—
√	√	54.6	53.4	53.0

%CER for dev02 using MLE trained systems and word-trigram LM

- HLDA used to project from static/1st/2nd/3rd derivatives to 39 dim
- Normalised pitch extracted using ESPS
  (+Pitch static/1st/2nd derivatives appended *after* HLDA)
- Results:
  - VTLN yields 1.5%-1.8% absolute reduction in CER
  - HLDA yields 0.8%-1.9% absolute reduction in CER
  - Pitch generally useful
- VTLN was used for all further experiments

Feature Vector Dimensionality

# Dim	HLDA		
	PLP	+Pitch	Pitch
39	53.8	—	53.4
42	53.7	53.0	53.2
45	53.8	53.3	53.1
48	—	53.4	53.1

%CER for dev02 using 12 mix comp MLE trained systems and word trigram LM

- Three systems examined:
  - PLP: baseline frontend with no pitch
  - +Pitch: baseline system with pitch added *after* HLDA
  - Pitch: HLDA projection from baseline frontend *and* pitch
- Small variation in performance with dimensionality
- Consistent gain ($\approx 0.5\%$) with using pitch in addition to HLDA

Additional Mixture Components/Varmix/MPE

# Comp	MLE	+Varmix	+MPE
12	53.0	52.2	49.8
16	52.3	51.7	49.9

%CER for dev02 using HLDA +Pitch trained systems and word trigram LM

- Varmix yields 0.6%-0.8% absolute reduction in error rate
- MPE yields 2.4% absolute for 12 component system
- 16 component system MLE systems better and MPE system about same
- Too many Guassians per hour for 16 comp MPE system!

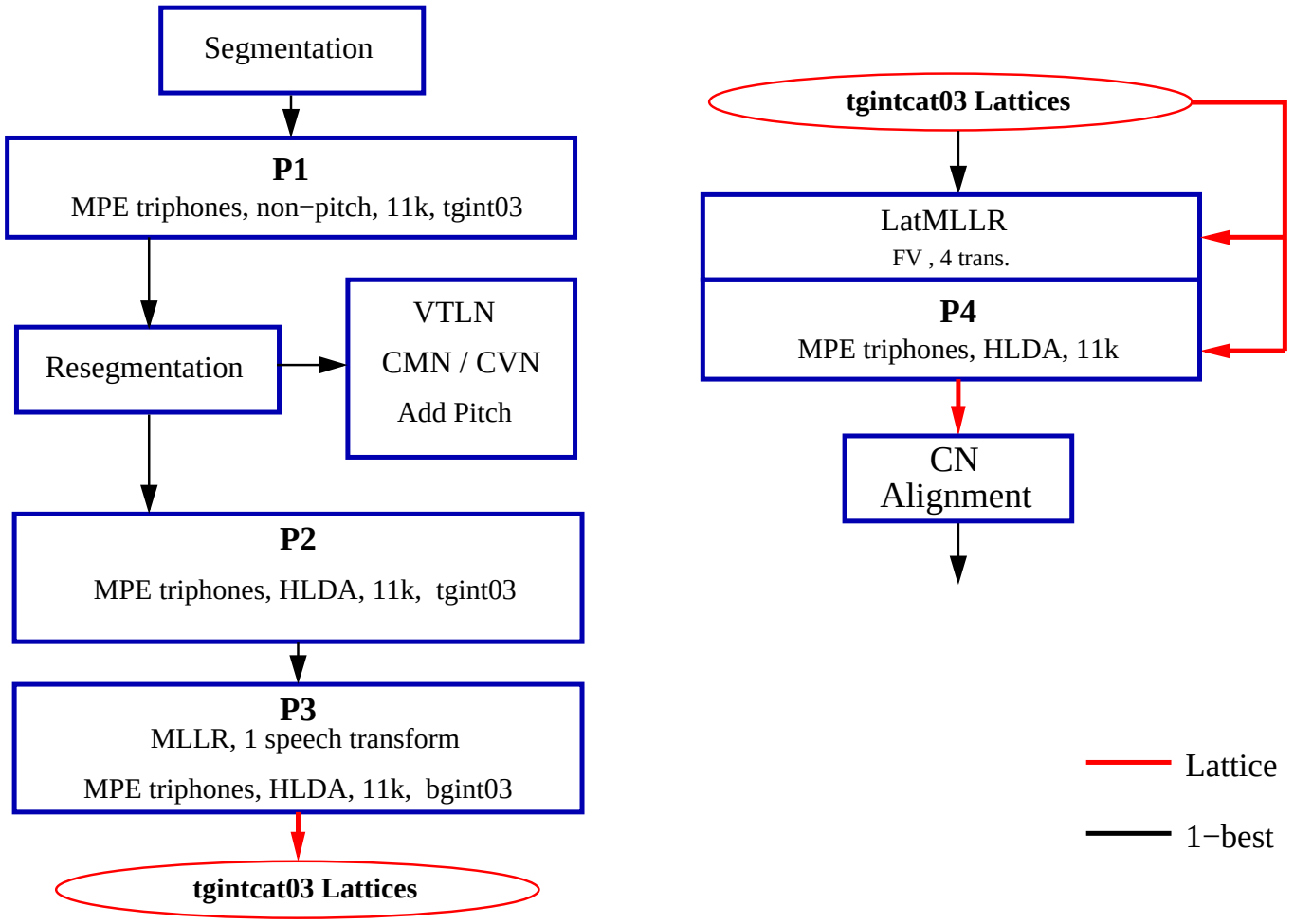
Automatic Segmentation

Segmentation	Diarisation			CER (%)
	MS	FA	Tot	
Manual	1.2	24.5	25.6	49.8
Automatic	3.7	8.3	11.9	50.8

%CER for dev02 using 12 mix comp HLDA +Pitch +Varmix MPE trained systems

- GMM classifier:
  - PLP with energy and channel energy difference plus 1st/2nd derivatives
  - 64 components for speech, 1024 components for silence.
- Diarisation score (% frame error) missed speech (MS), false alarm (FA):
  - reference derived from forced alignment of transcribed portions
  - untranscribed portions not scored (Manual MS score attribute of smoothing)
  - Manual segmentation error dominated by additional silence
- Automatic segmentation degraded CER by 1% absolute.

Mandarin RT03 System Overview



- Single system - currently no system combination.

Complete System Results

		CER (%)	
		dev02	eval03
P1	trans for VTLN	55.1	54.7
P2	trans for MLLR	50.8	51.3
P3	lat gen (bg)	49.3	50.5
	tgintcat rescore	48.9	49.8
P4	lat MLLR	48.6	49.5
CN	P4	47.9	48.6

%CER on dev02 and eval03 for all stages of 2003 system

- Final confidence scores have NCE 0.190 on eval03

Absolute Gains: dev02 vs eval03

Change to		ΔCER (%)	
		dev02	eval03
59-phone	46-phone	-1.1	-1.0
non-Tonal	Tonal	-1.3	-1.7
non-VTLN	VTLN	-1.8	-1.9
non-pitch	pitch	-1.1	-0.1
non-HLDA	HLDA	-1.9	-0.9

%CER changes on dev02 & eval03 using 12 comp MLE trained systems and word trigram LM

- dev02 numbers use manual segmentation, eval03 uses automatic segmentation
- Comparison of dev02 and eval03 gains:
  - all design choices gave improvements on both test sets
  - absolute gains differ (particularly pitch and HLDA), decisions affected by train/test speaker overlap?

Conclusions

- Current system:
  - 46 phone set, with tonal decision tree questions
  - 3 emitting states per phone model
  - VTLN, pitch, MPE and linear adaptation
  - standard techniques yield gains (but consistently less than expected)
- Future work:
  - investigate limited gains from standard schemes
  - additional systems, SAT etc, and system combination
  - alternative phone sets
  - modify HMM topology
  - add degree of voicing to frontend