DEVELOPMENT OF THE CUHTK 2004 RT04F MANDARIN CONVERSATIONAL TELEPHONE SPEECH TRANSCRIPTION SYSTEM

M.J.F. Gales, B. Jia, X. Liu, K.C. Sim, P.C. Woodland and K. Yu

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {mjfg,bj214,xl207,kcs23,pcw,ky219}@eng.cam.ac.uk

ABSTRACT

This paper describes the development of the CUHTK 2004 Mandarin conversational telephone speech transcription system. The paper details all aspects of the system, but concentrates on the development of the acoustic models. As there are significant differences between the available training corpora, both in terms of topics of conversation and accents, forms of data normalisation and adaptive training techniques are investigated. The baseline discriminatively trained acoustic models are compared to a system built with a Gaussianisation front-end, a speaker adaptively trained system and an adaptively trained structured precision matrix system. The final version of the evaluation system is then described. Results with an improved language model, using general web-data, are also presented.

1. INTRODUCTION

This paper presents the development of the CUHTK 2004 Mandarin conversational telephone speech transcription system. At Cambridge University HTK has been used to build large vocabulary speech recognition systems particularly for American English [1, 2]. In this work the techniques that have been developed for English transcription are applied to Mandarin conversational telephone speech recognition. However, since Mandarin is a tonal language, it is also necessary to change both the phone set and the acoustic front-end to incorporate information about tone.

The paper is organised as follows. In the next section the resources that were used are described, including the form of the phone set. The baseline acoustic model and front-end development are then described. This gives the baseline acoustic model that is used as a basis for the more advanced acoustic modelling techniques then discussed. The development framework and associated experimental results are presented along with the final evaluation system. Finally experiments with an improved language model, making use of general web-data [3], are described.

2. TRAINING DATA

This section briefly describes the resources and data that were used for the development of the Mandarin system.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors would also like to than the rest of the CUED EARS team for help in developing the Mandarin conversational telephone speech system. The authors would also like to thank the University of Washington for making the general web data available.

2.1. Dictionary and Phone Set

The original phone set and dictionary were supplied by the Linguistic Data Consortium (LDC). The dictionary consists of approximately 44,000 words and associated phonetic transcriptions. The LDC phone set consists of 60 phones and associated tone markers. It was found that one of the phones "u:e" occurred very rarely and so was mapped to "ue". This yielded a toneless phone set of 59 phones. In order to further reduce the number of phones, an additional mapping where long final phones were split was examined. Mappings of the form "[aeiu]n→[aeiu] n" were applied to the dictionary. This yielded a phone set of 46 phones. In initial experiments this 46 phone set was found to outperform the original LDC 59 phone set.

As Mandarin is a tonal language, incorporating the tone markers into the acoustic models should improve the system performance. Two ways of incorporating tonal information were investigated. The first used tonal phones as the basic phonetic unit for the decision trees. Alternatively, phonetic questions can be asked in the decision tree generation process. In preliminary experiments, there was little difference in performance between the two schemes, with both yielding gains over the toneless phone system. For this work tonal information was incorporated using the decision tree as this was felt to be more flexible and robustly handles the rare tonal phones. For all experiments the mapped 46 phone set and associated dictionary derived from the LDC dictionary were used with tonal decision tree questions. As there are no natural word boundaries in Mandarin, the characters may be partitioned into "words" in various ways. In this work the LDC character to word segmenter was used. This segmented data was used to generate the language model.

2.2. Acoustic Training and Test Data

The acoustic training data available for the 2004 CTS Mandarin task consists of two parts, ldc04 and swm03, yielding a total of about 72 hours of data. swm03 was made available for the 2003 RT04 Mandarin CTS task. It comprises two parts. The first section of 15.2 hours is part of the LDC CallHome data (chm). The second part is 16.6 hours of the LDC CallFriend data (cfm). ldc04 is a new data set for the 2004 system. It was collected by the Hong Kong University of Science and Technology (HKUST). There are 251 conversations (502 sides), corresponding to approximately 40 hours of training data. The test data for the 2004 evaluation was also collected by HKUST. Development data, dev04, was made available for this task comprising 2 hours of data, 24 conversations. The 2003 evaluation data, taken from the LDC CallFriend data, eval03, was also used to evaluate performance. This is a 1 hour test set of 12 conversations. However the primary development data was dev04, as this reflects the 2004 evaluation data.

2.3. Wordlists and Language Models

Source	# Words
ldc04	402.9K
swm03	401.7K
Mandarin TDT2	5.7M
Mandarin TDT3	4.0M
Mandarin TDT4	1.5M
People's Daily	70.3M
Xinhua	12.8M
China Radio	56.0M

Table 1. Number of words using the LDC character to word segmenter for each of the language model sources.

In addition to the 72 hours of acoustic training, six news corpora were used to train the language model, Mandarin TDT[2,3,4], China Radio, People's Daily and Xinhua. The size of each of the possible language model sources is shown in table 1. The LDC character to word segmenter was used for all sources. In order to determine the wordlists, all the words that occur in the acoustic training data were used. The two acoustic training data sources, and each of the news corpora, were kept as distinct sources for language model (LM) generation. Trigrams were generated for each of the sources and then interpolated.

Two sets of LMs are used in this work. The first set, tgint03 and tgintcat03, was built for the 2003 Mandarin system. As this LM was built prior to the availability of the ldc04 training, that acoustic data was not used. Thus the wordlist was only based on the swm03 training data and yielded an 11k wordlist. The interpolation weights were tuned on the eval03 test data[1]. As expected the interpolation weights were dominated by the acoustic training data, 0.88. The tgincat03 LM additionally used a class-based LM, 75 classes, built on the swm03 data. The second set of language models, tgint04, tgintcat04 and fgintcat04, was built with both the acoustic data sources and all the text corpora. Using all the words that appear in the acoustic training data gave an 17K wordlist. Again for interpolation the acoustic sources were heavily weighted. The differences in the topics was reflected in the fact that for example the word-based trigram language model, tgint04, the ldc04 LM component was weighted by 0.73 compared to the swm03 component with 0.15. The total contribution from all the news corpora was about 0.12, with the majority from People's Daily (0.09). A class based trigram language model was also built on the acoustic sources, using 200 classes estimated from all the acoustic training data. his was then interpolated with the word-based trigram language model, tgintcat04. The final language model, fgintcat04, interpolated four-gram language models from the acoustic sources and trigram language models from the news corpora. This interpolated model was then combined with the trigram class-based language model.

[1]Though the interpolation weights were tuned on the test data this has been found to yield no significant bias in the recognition results or perplexity, very few parameters are being estimated.

Language Model	eval03		dev04	
	PP	OOV	PP	OOV
tgint03	172.8	1.04	234.1	3.67
tgintcat03	160.4		280.8	
tgint04	218.4	0.50	173.2	1.03
tgintcat04	—		165.9	
fgintcat04	—		165.3	

Table 2. Perplexity (PP) and out of vocabulary (OOV) rates (excluding English words).

Table 2 shows the perplexity scores and the OOV rates[2]. The two sets of test data are clearly different. Using the 2003 language models, yields good perplexity scores on the the eval03 data, but poor scores on the dev04 data. The opposite is true for the 2004 language model. Using the interpolated class-based language model gave over a 7 point reduction in perplexity on dev04. A small additional gain was obtained using the four-gram model. As there is such a large difference between the two sets of data, the tgint04 language model will be used for all dev04 development results and the tgint03 language model for all eval03 development results. This allows the differences in performance of the various acoustic models to be concentrated on.

3. INITIAL DEVELOPMENT

This section describes the development of the baseline acoustic models. The initial models used only the ldc04 acoustic training data, as this is more closely related to the dev04 test data. A gender independent decision tree clustered triphone system was built with approximately 4,000 distinct states with 12 components per state. For testing a manually partitioned version of the dev04 test set was initially used (dev04PE) and an automatically segmented version of eval03 data (eval03).

3.1. Front-End Processing

The basic front-end for the Mandarin system was set to be similar to the English CTS system [1]. This uses a reduced bandwidth analysis, 125–3800 Hz, to generate 12 PLP Cepstra along with the zeroth Cepstra. First and second-order differences were appended to give 39 features. Cepstral mean and variance normalisation (CMN/CVN) was also applied per conversation side.

Training Data	Front-End	CER(%)
ldc04 (S1)	CMN/CVN	47.0
	+VTLN	43.2
	+HLDA	42.0
	+Pitch	41.6

Table 3. Baseline ML performance on dev04PE.

Table 3 shows the performance of the basic acoustic model with the baseline CMN and CVN front-end. This yielded an error

[2]The calculation of the OOV rates were based on the LDC character to word segmenter. Though the Mandarin OOV rate can be set to be zero by adding all single characters to the dictionary in preliminary experiments this made no difference to the CER.

rate of 47.0% on the dev04PE data. Using VTLN in both training and testing reduced the error rate by about 3.8% absolute. The front-end was then expanded to incorporate third-order differences and projected back to 39-dimensions using heteroscedastic LDA (HLDA) [4] and the efficient optimisation in [5]. This gave a further reduction in CER of 1.2%. It is also commmon for tonal languages to incorporate pitch into the front-end. Pitch was extracted using ESPS waves and normalised in a similar fashion to [6]. The pitch, along with the first and second-order differences, were then added after the HLDA projection[3], giving a complete feature vector of 42 dimensions. The final unadapted performance on the dev04PE test set was 41.6%.

After fixing the front-end, standard model building approaches used in the CUED evaluation systems were applied. The number of components per state was made proportional to the amount of training data for that state, though keeping the average number the same, and minimum phone error (MPE) training applied [7]. For the MPE training a modified version of the dynamic maximum mutual information (MMI) prior presented in [8]. Here two levels of I-smoothing were used. The first level smoothed the MMI statistics with the ML statistics. The second level smoothed the MPE statistics with the smoothed MMI statistics. This was found to be more robust that the direct use of the MMI statistics as there is only 72 hours of training data. The final error rate after MPE training was 38.2% on dev04PE.

3.2. Model Structure

This section describes the initial development of the acoustic models. For this work both the dev04 and eval03 test sets were used for development. The tgint04 LM was used for the dev04 test set and the tgint03 LM was used for the eval03 test set. This was felt to be necessary because of the difference in topics illustrated by the large differences in the perplexity scores shown in table 2.

Training Data		Avg. Comp.	CER(%) dev04PE	eval03
swm03	—	12	—	48.6
ldc04	S1		38.2	56.6
	S2	12	36.3	48.2
ldc04+swm03	S3	16	36.1	47.9
	S4	20	36.0	47.2

Table 4. Baseline MPE model performance.

Table 4 shows the performance of various MPE trained systems. The first line, swm03, was trained using only the 2003 swm03 training data. This is simply to show a baseline number on eval03. It is clear that in addition to the differences in topic, there are also accent, possibly channel, differences between the 2003 and 2004 data sets. For the ldc04 trained system the performance on eval03 was 8.0% absolute worse than that of the swm03 trained system.

ldc04 and swm03 were then combined together, though keeping the decision tree and HLDA projection from the ldc04 data. This is the S2 system in table 4. Not surprisingly using the 2003 training data significantly reduced the error rate on the eval03 test data. The performance of the S2 system is better than the swm03 trained system. In addition the error rate on the dev04PE test set was also improved, though to a lesser extent than the eval03 data. With the additional training data, additional components may be robustly trained. Using an average of 16 components, the S3 system, gave an additional 0.2% on dev04PE. An additional 4 components, the S4 system, gave minimal difference on dev04PE, but did decrease the error rate on the eval03 data. Since the primary test was the dev04 data, the S3 system was selected as the starting point for further comparisons.

Decision Tree/HLDA generation data		CER(%) dev04PE	eval03
ldc04	S3	36.1	47.9
ldc04+swm03	S5	36.4	47.2

Table 5. Performance varying the decision tree and HLDA training data, all models MPE trained.

All the ldc04 and ldc04+swm03 trained systems shown in table 4 used the same decision tree and HLDA projection. Table 5 shows a comparison of the S3 system with training the decision tree and HLDA projection on all the training data, rather than just the ldc04 data. The effects of tuning the projection and decision tree to a particular task are clear. Training a tree and projection on all the data yielded lower error rates on the eval03 data, but higher error rates on the dev04 data than the S3 system.

3.3. Segmentation

For the actual evaluation the segmentation for each side of the conversation is not given. In order to segment the data a simple GMM classifier was used. The features used in the automatic segmentation were 12 PLP features, log energy, energy difference between channels, and corresponding delta and acceleration coefficients. Three GMMs (male speech, female speech and silence) were used to identify speech segments. Both speech models had 64 Gaussian components and the silence model had 1024 components.

Segmentation	Diarisation Scores MS	FA	DER	CER (%)
Manual (dev04PE)	—	—	—	36.1
Automatic (dev04)	3.6	5.8	9.42	37.3

Table 6. Effect on dev04 performance using manual versus automatic segmentation with the S3 MPE unadapted acoustic model, including diarisation results for missed speech (MS), false alarm (FA) and diarisation error rate (DER).

Table 6 shows the effect of the use of an automatic segmenter on the dev04 test data. The MPE trained S3 system was run on the automatically segmented data. The increase in error rate from using the automatic segmentation was about 1.2% absolute[4].

[3] In initial experiments there was little difference between using HLDA on the complete feature vector and projecting just the PLP features. As the final P1 model is a non-Pitch model, using an HLDA projection of just the PLPs simplifies the system build.

[4] The DER are slightly different to those in the submitted ICASSP paper as the the previous results did not use the official UEM file.

4. ACOUSTIC MODELS

In the previous section the development of the baseline acoustic models was described. For the 2004 CTS English system [9] a variety of more advanced acoustic models were investigated. This section briefly describes some of these models. As in the English CTS development these advanced models were used to rescore lattices generated within the 10xRT framework described in section 5.1 The three corpora, chm, cfm and ldc04, differ in terms of dominant accents and topics. It is therefore useful to examine the forms of normalisation for the data.

4.1. Gaussianisation

The use of CMN and CVN transforms the feature vector so that the mean of each dimension for each side is 0 and the variance is 1. There is no matching of the higher-order statistics. Histogram normalisation is one approach that has been used to further normalise data for CTS-English on a per-speaker basis [10]. A modified version of this using a smoothed form based on a per-dimension GMM is used in this work. It was found that there was little difference in performance between the histogram approach and the use of GMMs, however the GMM yields a more compact, smoother estimate, of the histogram. The feature-vector transformation for element i of the observation $\mathbf{o}$, o_i, is

$$\tilde{o}_i = \phi^{-1}\left(\int_{-\infty}^{o_i} \sum_{m=1}^{M} c_i^{(sm)} \mathcal{N}(x; \mu_i^{(sm)}, \sigma_i^{(sm)2}) dx\right) \quad (1)$$

where $\phi^{-1}()$ is the standard Gaussian inverse cumulative density function, $c_i^{(sm)}$, $\mu_i^{(sm)}$ and $\sigma_i^{(sm)2}$ are the prior, mean and variance for the i^{th} dimension for side s of component m. The components of the GMM are trained on a per-side basis, indicated by s, after the application of the HLDA projection. All elements of the feature vector, including pitch, were normalised.

4.2. Speaker Adaptive Training

An alternative approach to normalising the features is to use a linear transformation. One standard approach is to use constrained MLLR, where the linear feature transformation is estimated by maximising the likelihood of the data [11]. This is speaker adaptive training (SAT). The form of the transformation for vector $\mathbf{o}$ is

$$\tilde{\mathbf{o}} = \mathbf{A}^{(s)}\mathbf{o} + \mathbf{b}^{(s)} \quad (2)$$

The linear transformation parameters, $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ are trained for each side s. One of the disadvantages of SAT is that in order to estimate the test speaker transformation either supervised adaptation data, or some initial hypothesis, is required. This is not the case for Gaussianisation as a GMM is simply estimated on all the data from one-side. To ease this problem a corpus-based form of adaptive training was examined. Here a linear transform was estimated for each corpus and the models adaptively trained. However, this only gave slight improvements in performance, approximately 0.2%, over the baseline HLDA system in the development framework (table 7) so was not considered further.

4.3. Structured Precision Matrices

The baseline acoustic models are based on states with output distributions using GMMs with diagonal covariance matrices. Recently there has been work on using structured forms of precision matrix models. The form of model used in this paper is based on SPAM [12, 13]. Here the precision, inverse covariance, matrix can be written as

$$\mathbf{\Sigma}^{(m)-1} = \sum_{i=1}^{R} \lambda_i^{(m)} \mathbf{S}^{(i)} \quad (3)$$

where $\lambda_i^{(m)}$ are the basis co-efficients for each component in the system m and $\mathbf{S}^{(i)}$ is the i^{th} basis matrix. For this work R was set to be 39. For details of the basis matrix initialisation and MPE training of these models see [13]. As there is significant variability in the training corpora a SAT-SPAM system was built, where a discriminative SPAM system was estimated in the space defined by a SAT trained system [14].

5. DEVELOPMENT RESULTS

The system used for the experiments was based on the 2003 CUHTK CTS English Rich Transcription evaluation system. A multi-branch, multi-pass approach is used along with system-combination. For details of the English versions of this framework see [1].

5.1. Development Framework

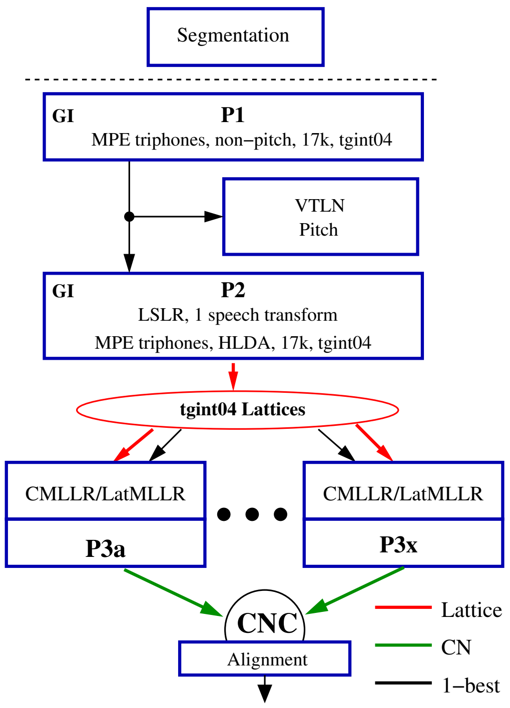

Fig. 1. System structure (note tgint03 LM used for eval03).

Figure 1 shows the basic structure of the system. P1 is used to provide an initial transcription for VTLN estimation. After VTLN estimation, pitch is added to the features and the P2 models are adapted using least squares regression mean and diagonal variance transforms to the P1 hypothesis. This adapted P2 model is then used to generate lattices for rescoring in the P3 stage. For the P3 stage, all models are adapted using speech and silence constrained MLLR transforms and the P2 hypothesis. They are then further adapted using lattice MLLR to estimate mean and diagonal variance transforms.

The final system output was derived by combining the confusion networks generated by the P3a to P3x passes using Confusion Network Combination (CNC) [15]. Finally, a forced alignment of the final word-level output was used to obtain accurate word times before scoring. For this initial development work, no note was taken of the run-times, though the evaluation has real-time constraints.

System	S3	CER (%)	
		dev04	eval03
P2-tgint04/03	HLDA	37.1	46.9
P2-cn		36.2	44.9
P3a-cn	HLDA	35.8	45.0
P3b-cn	SAT	35.0	44.2
P3s-cn	SAT-SPAM	34.2	43.7
P3d-cn	GAUSS	34.6	43.3
P3f-cn	GAUSS-STC	34.2	43.0
P3g-cn	GAUSS-SPAM	33.5	42.4
P3e-cn	GAUSS-SAT	33.8	42.6
P3t-cn	GAUSS-SAT-SPAM	33.2	41.8
P3d+P3s	CNC	33.4	42.9
P3g+P3t		33.1	41.8
P3e+P3t		32.8	41.4

Table 7. Development CER on dev04 and eval03.

Table 7 shows the results for the acoustic models within the development framework. All the P3 numbers are given after confusion network (CN) decoding. The performance of the baseline MPE model (HLDA) in the P3 stage was disappointing compared to the SAT system. Using SAT the error rate on both dev04 and eval03 was decreased by 0.8% absolute compared to the HLDA system. This shows the large variability of the acoustic training data. This error rate was further decreased using the SAT-SPAM system, which had the lowest CER on the dev04 data, 34.2%.

As there is such large variability in the acoustic data, additional normalisation should be helpful. Gaussianisation was therefore used to further normalise the data (GAUSS). From table 7, this reduced the error rate by 1.2% and 1.7% absolute on the dev04 and eval03 tasks compared to the HLDA system. Since Gaussianisation is a non-linear transformation of the features, it is useful to apply an additional level of a global semi-tied transformation [5] to the Gaussianised features[5] (GAUSS-STC). This gave an addition 0.3 to 0.4% absolute reduction in error rate. Rather than using as semi-tied transform, SPAM could be used in the Gaussianised space (GAUSS-SPAM). This gave a further 0.6% to 0.7% reduction in error rate over the semi-tied system. Further normalisation of the features can be obtained by combining the Gaussianisation system with the CMLLR adaptive training. This gave the lowest error rate with the GAUSS-SAT-SPAM system giving 33.2% and 41.8% on dev04 and eval03 respectively. This is 2.6% and 3.2% absolute better than the standard HLDA system.

[5]For the standard HLDA transform this should yield little difference in performances the HLDA transform can subsume the semi-tied transform

5.2. Refinements

The development framework was simpler than the standard CUHTK evaluation systems. Various refinements to this basic development framework are described in this section.

5.2.1. Full Language Model

System (S3)	LM	CER (%)
		dev04
P2 (HLDA)	tgint04	37.1
	tgintcat04	36.7
	fgintcat04	36.6

Table 8. Lattice expansion using class based LM with CER on dev04.

The language model used in section 5.1 was a trigram language model. From table 2 the perplexity of the data can be reduced by using class-based language models and four-grams. The trigram P2 lattices from table 7 were expanded using the interpolated class-based and word-based trigram language model (tgintcat04) and the interpolated class-based trigram and word-based four-gram (fgintcat04). Table 8 shows the effect this has on the CER for dev04. Using the more complex language models reduced the CER, with the class-based LM giving about 0.4% absolute and the four-gram a small gain of about 0.1%.

5.2.2. Lattice Regeneration and Task Porting

In previous systems [2] the use of *lattice regeneration* and *lattice combination* have yielded gains. Here, part way through the MPE training process a new set of lattices are generated using the latest MPE trained model. Typically this is done after 4 iterations. The results shown in table 7 shows that considerable normalisation is required to compensate for the acoustic and speaker differences within and between the two databases. Given these observations it is sensible to also examine task discriminative porting [16] to the ldc04 data.

System (S3)	CER (%)
	dev04
GAUSS	36.2
+regen/comb	36.0
+ldc04 port	35.9

Table 9. Unadapted decode numbers adding lattice regeneration and combination and ldc04 task porting to the MPE-GAUSS system.

Table 9 shows the performance of task porting to the ldc04 data after using lattice regeneration on top of a system with a Gaussianisation front-end. Note for the unadapted numbers the Gaussianised frontend gave about 1.1% absolute over the HLDA numbers in table 6. The use of lattice regeneration and combination[6] gave about 0.2% absolute over the standard MPE system. The class porting gave a further 0.1% absolute. The use of gender-dependent models as used in the English CTS system gave similar gains to the task porting. However using the task porting required no gender labelling of the conversation sides.

5.2.3. Improved P2 and Adaptation

From the results in the previous section the use of Gaussianisation, which does not require any information about the word sequence, yields over a 1.0% reduction in CER. This is therefore a good candidate to replace the initial HLDA model set for lattice generation. The adaptation used in the development framework was based on lattice-MLLR. An alternative approach is to use confidence-based MLLR [17] where the accumulates are weighted by the confidence of each word in the hypothesis.

System	S3	Adapt	CER (%)
			dev04
P2-tgint04	GAUSS		34.8
P2-fgintcat04			34.2
P2-cn			33.8
P3e-l	GAUSS-SAT	(lattice)	33.0
P3e-c		(conf)	32.8
P3t	GAUSS-SAT-SPAM	(lattice)	32.1
P3e-l+P3t	CNC		31.9
P3e-c+P3t			31.8

Table 10. Lattice-based MLLR (lattice) versus confidence-based MLLR (conf). P2 lattices generated using GAUSS system and expanded using fgintcat04.

For these experiments a GAUSS system was used to generate the P2 lattices and the fgintcat04 LM used during the lattice expansion. From table 10 this gave a reduction in P2 error rate of 2.3% absolute (using tgint04) over the HLDA system. For the final system combination stage, using lattice combination, the error rate was reduced by by 0.9% absolute by rescoring GAUSS generated lattices rather than HLDA generated lattices and using the more complex language model. Table 10 also shows a comparison of the use of confidence-based versus lattice-based MLLR for the GAUSS-SAT system. The use of confidence-based MLLR gave about 0.2% absolute in the single branch and 0.1% absolute in the final combination.

6. EVALUATION SYSTEM

The final evaluation system was based on the development framework combined with the improvements in the previous section and tuning of the P2 parameters to yield a less than 10xRT system (10xRT) and less than 20xRT system (20xRT). The following acoustic and language models were used at each of the stages. All models used the same decision tree and HLDA projection trained on the ldc04 data and an average of 16 components per state (the S3 system).

- **P1**: an MPE-trained non-VTLN HLDA system using the ldc04+swm03 training data with the tgint04 language model.
- **P2**: an MPE trained GAUSS system with lattice regeneration and task porting to the ldc04 data. The lattice expansion for subsequent P3x rescoring used the fgintcat04 language model.
- **P3e**: an MPE-trained GAUSS-SAT system. No lattice regeneration or task porting was used[7]. This branch was adapted using confidence-based MLLR with word confidence scores from P2-cn.
- **P3t**: an MPE-trained GAUSS-SAT-SPAM system with lattice regeneration and task porting. This branch was adapted using lattice-based MLLR.

In addition the segmentation was slightly modified. Though the unadapted error rates for the segmentation were the same the deletion rates were found to be slightly lower using the modified segmentation. The eval04 test set consists of about 1 hour of data taken from 12 conversations (24 sides).

6.1. Results

System	S3	CER (%)	
		10xRT	20xRT
P1	HLDA (non-VTLN)	43.6	43.6
P2-tgint04	GAUSS	34.3	34.3
P2-fgintcat04		34.1	34.1
P2-cn		33.9	33.7
P3e-cn	GAUSS-SAT (conf)	33.1	32.8
P3t-cn	GAUSS-SAT-SPAM	32.1	32.1
P3e+P3t	CNC	32.1	31.7
P3e+P3t+P2		32.0	32.0

Table 11. Final 10xRT and 20xRT CER dev04 performance.

Table 11 shows the development results using the final version of the evaluation system. For the final system combination, in addition to the two P3 branches, it is possible to use the P2-cn output. For the 10xRT system this was found to give a slight gain in performance, whereas for the 20xRT system a slight degradation was obtained. P2 was therefore only combined in the 10xRT system.

System	S3	CER (%)	
		10xRT	20xRT
P2-cn	GAUSS	31.9	31.8
P3e-cn	GAUSS-SAT (conf)	30.6	30.5
P3t-cn	GAUSS-SAT-SPAM	30.0	30.2
P3e+P3t	CNC	29.4	29.5†
P3e+P3t+P2		29.7†	29.7

Table 12. Final 10xRT and 20xRT CER eval04 performance. † indicates the system submitted for the evaluation

[6]For these experiments the lattice regeneration was performed using an HLDA system, where a gain of 0.3% absolute over the standard MPE trained system was obtained.

[7]A slight degradation in performance was found when lattice regeneration and task porting were used.

Table 12 shows the breakdown of the evaluation system performance on the eval04 test set. Overall similar trends to the development numbers shown in table 11 can be seen. The one major difference is that the 10xRT performance was degraded by the use of the P2 branch in the CNC stage. Strangely the 10xRT performance using the two-way P3 CNC system was slightly better than the 20xRT system. The final primary submitted system gave error rates on eval04 of 29.7% and 29.5% for the 10xRT and 20xRT systems respectively.

6.2. Processing Speed

Test Set	RT factor	
	10xRT	20xRT
dev04	8.2	12.4
eval04	8.9	13.3

Table 13. dev04 and eval04 evaluation system real-time factors.

Table 13 shows the breakdown of the real-time factors for the evaluation system on the dev04 and eval04 test sets. These were obtained using a single CPU IBM x335 machine (3.2GHz Intel Xeon, 2MB L3 cache 533MHz Bus) in hyperthreading mode running Linux.

7. GENERAL WEB DATA LANGUAGE MODEL

After the evaluation the University of Washington made available additional Mandarin language model training data from the web. The data was collected using Google with frequent N-gram queries. Further details of the methodology used to collect this *general web-data* (GWD) is described in [3]. The data consists of approximately 115 million words (about 102K unique words of which 58K were English "words"). A word-based trigram language model was generated using this GWD training data and used as an additional possible source for interpolation.

Source	# Words	Language Model	
		tgint04	+gwd
ldc04	402.9K	0.73	0.62
swm03	401.7K	0.15	0.10
People's Daily	70.3M	0.09	0.01
GWD	115.0M	—	0.26
others	—	0.03	0.01

Table 14. Interpolation weights for the word-based trigram language model with and without the general web data (GWD).

In the same fashion as the language models in table 2 (and the ones used in the evaluation) the interpolation weights were estimated using dev04. Table 14 shows the interpolation weights for the various sources for the word-trigram language models. A significant weight, 0.26, is applied to the general web data (GWD). Similar trends were observed when four-gram acoustic data sources were used. The interpolation weights between the four-gram acoustic model LMs and trigram news sources with the class-based trigram language model were 0.78 and 0.22 respectively.

Language Model	dev04		eval04	
	—	+gwd	—	+gwd
tgint04	173.2	156.0	166.2	144.4
tgintcat04	165.9	151.8	160.7	142.8
fgintcat04	165.3	150.0	159.9	140.9

Table 15. Perplexity for original LM and LM including the general web data (+gwd). OOV rate for eval04 was 0.75% excluding English words.

Table 15 shows the effect of using the GWD on the perplexity of the dev04 and eval04 data. Note the OOV rate for the eval04 data was 0.75% exluding English words, similar to the 1.03% for dev04. The use of the GWD gave significant reductions in the perplexity. On the eval04 data an almost 20 point reduction in perplexity was obtained. In [3] the news corpora were not used in the interpolated language model. It was found that the additional news corpora gave a slight reduction in perplexity compared to not using them, for tgint04 not using the news-sources gave 156.9 compared to 156.0 including the news corpora for dev04 and 145.3 and 144.4 for eval04. All experiments use the interpolated language model with the news corpora.

Language Model	System (S3)	CER (%)
		dev04
tgint04	GAUSS	36.2
tgint04+gwd		35.1

Table 16. Unadapted performance of MPE Gaussianised system on the dev04 test set using the original LM (tgint04) and new LM with global web-data (tgint04+gwd).

The CER performance of the new word-tri-gram language models with the general web-data (tgint04+gwd) was initially evaluated using unadapted S3 GAUSS models. The new LM gave 1.1% absolute reduction in CER on the dev04 test set.

Test Set	CNC	10xRT		20xRT	
		—	+gwd	—	+gwd
dev04	P3e+P3t	32.1	30.8	31.7	30.6
	P3e+P3t+P2	32.0	31.0	32.0	30.9
eval04	P3e+P3t	29.4	28.6*	29.5†	28.3*
	P3e+P3t+P2	29.7†	28.6	29.7	28.5

Table 17. 10xRT and 20xRT CER dev04 and eval04 performance using language models with and without the general web data (+gwd). † indicates the primary systems submitted and $*$ indicates the late contrast systems submitted.

The new language models were then used with the evaluation framework of section 6. Initial decoding for P1 and P2 used the tgint04+gwd word-based trigram language model. The lattices were then expanded using fgintcat04+gwd. Table 17 shows the final performance at the confusion network combination (CNC) stage of the evaluation system on the dev04 and eval04 test sets. Within the evaluation framework use of the GWD in the language gave between 0.8% and 1.2% absolute gain reduction in CER. The best performance for the eval04 task was 28.3% CER. This is 1.2% absolute better the system without the GWD.

8. CONCLUSIONS

This paper has described the development of the CUHTK 2004 Mandarin conversational speech transcription system. The paper has concentrated on the possible forms of acoustic model that could be used. In particular, as there are significant differences in the acoustic training data, two forms of data normalisation were investigated, Gaussianisation and speaker adaptive training. In addition, the use of structured precision matrices was investigated, Results were presented within a multi-pass system combination framework, similar to the 2003 CTS 10xRT framework. Both forms of normalisation and the use of structured transforms reduced the character error rate. The use of Gaussianisation in combination with SAT and SPAM yielded the lowest CER. The final CER on the eval04 using the submitted system was 29.7% and 29.5% for the less than 10xRT and 20xRT systems respectively. After the evaluation a new language model was built using the general web-data from the University of Washington. This gave a reduction in error rate of around 1.0% absolute over the original evaluation system. The performance of this system, submitted as a lated contrast, was 28.6% and 28.3% for the less than 10xRT and 20xRT systems respectively.

9. REFERENCES

[1] G. Evermann, H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P.C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proc. ICASSP*, 2004.

[2] D.Y. Kim, G. Evermann, T. Hain, D. Mrva, L. Tranter, S.and Wang, and P.C. Woodland, "Recent advances in broadcast news transcription," in *ASRU*, 2004.

[3] T Ng, M Ostendof, M-Y Hwang, M Siu, I Bulyko, and X Lei, "Web-data augemnted language models for mandarin conversational speech recogntion," in *Submitted to ICASSP*, 2005.

[4] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, Johns Hopkins University, 1997.

[5] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[6] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Proc. Eurospeech*, 1997.

[7] D. Povey and P.C. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002.

[8] G Saon, D Povey, and G Zweig, "CTS decoding improvements at ibm," Presented at EARS STT Workshop, St. Thomas 2003., 2003.

[9] X. Liu, M.J.F. Gales, K.C. Sim, and K. Yu, "Investigation of acoustic modeling techniques for LVCSR systems," in *Submitted to ICASSP 2005*, 2005.

[10] G. Saon, A. Dharanipragada, and D. Povey, "Feature-space Gaussianization," in *Proc. ICASSP*, 2004.

[11] M J F Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[12] S. Axelrod, R. Gopinath, and P Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. ICSLP*, 2002.

[13] K C Sim and M J F Gales, "Precision matrix modelling for large vocabulary continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR485, Cambridge University, 2004.

[14] K.C. Sim and M.J.F. Gales, "Adaptation of precision matric models on large vocabulary continuous speech recognition," in *Submitted to ICASSP*, 2005.

[15] G. Evermann and P.C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings Speech Transcription Workshop*, 2000.

[16] D. Povey, P.C. Woodland, and M.J.F. Gales, "Discriminative MAP for acoustic model adaptation," in *Proc. ICASSP*, 2003.

[17] L F Uebel and Woodland P C, "Improvements in linear transforms based speaker adaptation," in *Proceedings ICASSP*, 2001.