DEVELOPMENT OF THE CUHTK 2004 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 models are finally evaluated within a multi-pass, mult-branch, system combination framework.

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. 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. Finally the development framework and some experimental results are presented.

2. TRAINING DATA

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

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. 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.

Acoustic Training and Test Data: The 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.

Word-lists and Language Models: 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. In order to determine the word-lists, 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 two, tgint03

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.

and tgintcat03, were 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 word-list was only based on the swm03 training data and yielded an 11k word-list. 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 built on the swm03 data. The second language model, tgint04, 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 16K word-list. Again for interpolation the acoustic sources were heavily weighted. The differences in the topics was reflected in the fact that 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). In contrast to the 2003 LM a class-based language model was not generated.

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

Table 1. Perplexity (PP) and out of vocabulary (OOV) rates (excluding English words).

Table 1 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. As there is such a large difference between the two sets of data, the tgint04 LM will be used for all dev04 development results and the tgint03 LM 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).

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 2. Baseline ML performance on dev04PE.

Table 2 shows the performance of the basic acoustic model with the baseline CMN and CVN front-end. This yielded an error 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 heteroscadastic LDA (HLDA). This gave a further reduction in CER of 1.2%. It is also common for tonal languages to incorporate pitch into the front-end. Pitch was extracted using ESPS waves and normalised in a similar fashion to [2]. 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 [3].

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 1.

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

Table 3. Baseline MPE model performance.

Table 3 shows the performance of various MPE trained systems. The first line, swm03, was trained using 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.

[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.

[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.

[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.

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 3. 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 16 components, the S3 system, gave an additional 0.2% absolute reduction in CER 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 4. Performance varying the decision tree and HLDA training data, all models MPE trained.

All the ldc04 and ldc04+swm03 trained systems shown in table 3 used the same decision tree and HLDA projection. Table 4 shows a comparison of the S3 system with training the decision tree and HLDA projection on all the training 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.

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. Table 5 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.

Segmentation	Diarisation Scores			CER
	MS	FA	DER	(%)
Manual (dev04PE)	—	—	—	36.1
Automatic (dev04)	3.6	5.7	9.3	37.3

Table 5. 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).

4. ACOUSTIC MODELS

In the previous section the development of the baseline acoustic models was described. For the 2004 CTS English system [4] 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

As the three corpora, chm, cfm and ldc04, differ in terms of dominant accents and topics, it is particularly useful to examine forms of normalisation for this training data.

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 [5]. 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.

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. 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 system, so was not considered further.

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 [6, 7]. 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 [7]. 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 [8].

5. DEVELOPMENT FRAMEWORK

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].

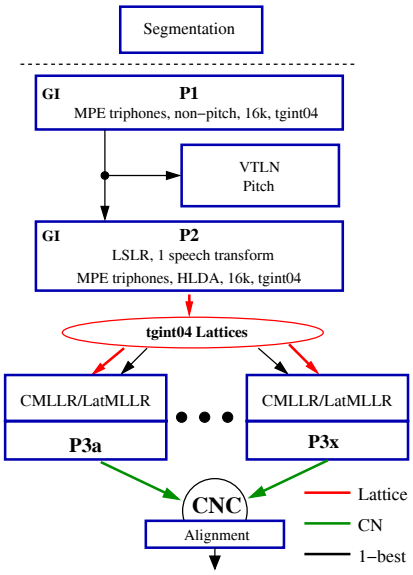

Fig. 1. System structure (note tgint03 LM used for eval03).

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 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). 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.

Table 6 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. This shows the large variability of the acoustic training data. This error rate was further decreased using the SAT-SPAM system to 34.2% on dev04. The use of the Gaussianisation front-end further improved the performance of all the systems by about 1% absolute on dev04. The best single system, 33.2% on dev04, was obtained using HLDA with Gaussianisation, SAT and SPAM covariance modelling. Combining the GAUSS-SAT-SPAM system with a GAUSS-SAT system gave an additional 0.4% absolute reduction in CER, to give 32.8% CER on dev04.

System	S3	CER (%)	
		dev04	eval03
P2	HLDA	37.1	46.9
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
P3e-cn	GAUSS-SAT	33.8	42.6
P3t-cn	GAUSS-SAT-SPAM	33.2	41.8
P3e+P3t	CNC	32.8	41.4

Table 6. CER on dev04 and eval03.

6. 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. In contrast to the English CTS system [4], the use of normalisation techniques, especially Gaussianisation, gave significant gains over the standard HLDA front-end. This is felt to be at least partly due to the greater variability and reduced size of the available training data rather than an inherent attribute of Mandarin. The final evaluation system using these acoustic modelling approaches achieved one of the lowest CER on the evaluation task.

7. 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] 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.

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

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

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

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

[7] 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.

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