THE CU-HTK MANDARIN BROADCAST NEWS TRANSCRIPTION SYSTEM

R. Sinha, M.J.F. Gales, D.Y. Kim, X.A. Liu, K.C. Sim and P.C. Woodland*

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {rs460,mjfg,dyk21,xl207,kcs23,pcw}@eng.cam.ac.uk

ABSTRACT

This paper discusses the development of the CU-HTK Mandarin Broadcast News (BN) transcription system. The Mandarin BN task includes a significant amount of English data. Hence techniques have been investigated to allow the same system to handle both Mandarin and English by augmenting the Mandarin training sets with English acoustic and language model training data. A range of acoustic models were built including models based on Gaussianised features, speaker adaptive training and feature-space MPE. A multi-branch system architecture is described in which multiple acoustic model types, alternate phone sets and segmentations can be used in a system combination framework to generate the final output. The final system shows state-of-the-art performance over a range of test sets.

1. INTRODUCTION

This paper presents the development of the CU-HTK Mandarin Broadcast News (BN) transcription system. The basic system shares features from the CU-HTK Mandarin conversational telephone speech (CTS)[1] system. However, for Mandarin BN it is also necessary to be able to deal with a significant amount of English speech data contained in the broadcasts. One approach to this issue, described in [2], uses a language identification stage to tag the English speech data. This approach was not found to perform reliably across different types of test data. In this paper the construction of acoustic and language models able to recognise both Mandarin and English data is investigated.

A range of development results are given. From a standard baseline system, the effects of incorporating English training data into both acoustic and language models are described. In addition results using language model data collected from the web are also described. The final Mandarin BN system described uses a multi-pass multi-branch approach in a system combination framework. Diversity in the system was introduced using multiple segmentations, two alternate phone sets and multiple acoustic model types. The acoustic models used for system combination were trained using the minimum phone error (MPE) criterion with Gaussianised features and can include speaker adaptive training and feature-space MPE (fMPE).

2. BASELINE SYSTEM

Acoustic Training and Test data: A total of 148 hours of data was available for acoustic model training. This comprises 28 hours of Hub-4 data released by the Linguistic Data Consortium (LDC) with accurate transcriptions. For the remaining 120 hours of TDT4 Mandarin BN data only closed-captions references were provided, hence light supervision [3] techniques were used this data. Of this 148 hours of data, approximately 1 hour of the TDT4 data comprised English. Two test sets were used for system development. The first was the RT-04 development data and consists of a total of 0.5 hours of CCTV data from shows broadcast in November 2003 (`dev04f`). The second set was the mainland shows (CCTV, VOA, and CNR) from the RT-03 evaluation containing 0.6 hours of data from February 2001 (`eval03m`). The system was also tested on the RT04 evaluation test set which includes a total of 1 hour of data from CCTV, RFA and NTDTV broadcast in April 2004 (`eval04`).

Language Model and Word-List: One issue in language modelling for the Chinese language is that there are no natural word boundaries in normal texts. A string of characters may be partitioned into "words" in a range of ways. There are multiple valid partitions that may be used. As in the CU-HTK Mandarin CTS system [1], for BN Mandarin texts the LDC character to word segmenter was used. This implements a simple deepest first search method to determine the word boundaries. Any Chinese characters that are not present in the word list will be treated as individual words. As part of this process a word-list is required. The basis for the Mandarin word-list was the 44K LDC Mandarin word-list. As English words were present in the acoustic training data transcriptions, all English words and single character Mandarin words not in the LDC list were added to the word-list to yield a total of 50K words. This is the word-list used for initial development.

The language model was trained using 366M words from five sources all released by LDC: the correct acoustic transcripts for Hub4 Mandarin data, China Radio, Mandarin TDT[2,3,4], Gigaword (Xin Hua) and People's Daily. During language model training the two acoustic sources, Hub4 and TDT4 Mandarin, and each of the news corpora, were kept as distinct sources. Word based trigram and 4-gram language models were generated for each of the sources and then interpolated. This is referred to as `lm1.0` in this paper and used in the initial system development.

Dictionary and Phone set: The baseline system was built using the 50K word-list described above for both language and acoustic model training and testing. This 50K word-list covers all English and Mandarin words in the acoustic training data. Two phone sets were used to build the dictionary. Both were derived from the 60 toneless phones used in the LDC 44K dictionary. The first phone set is the same as the one used in CU-HTK 2004 Mandarin CTS system [1]. It consists of 46 toneless phones, obtained by applying mappings of the form "[aeiu]n→[aeiu] n", "[aeiou]ng→[aeiou] ng" and "u:e→ue" to the LDC phone set. In addition a more compact phone set containing 38 toneless phones was derived by further splitting longer vowels as, for example, "u[ao]→u [ao]", "i[aou]→i [aou]" and "uai→u ai". For both phone-sets pronunciations for Mandarin characters not in the

Do Yeong Kim is now with VoiceSignal Technologies.

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.

44K LDC word-list were added manually. Automatic mapping rules from the CU-HTK English phone set to each of the Mandarin phone-sets was used for all the English words. Unless otherwise stated, the 46 phone set was used for all the experiments in the following sections.

Front-End Processing: The front-end for the Mandarin BN system used a 25ms frame-size with a 10ms frame-rate. Each frame was coded using 12 PLP coefficients along with zeroth cepstra with first, second and third derivatives appended and then projected down to 39 dimensions using heteroscadastic LDA (HLDA). Pitch, along with first and second derivatives, were then added to yield a 42 dimensional feature vector. This will be referred to as the HLDA front-end in this paper. Gaussianisation was found to be useful in the Mandarin CTS system [1]. Hence a Gaussianised front-end, GAUSS, was also investigated in this paper. As in the CTS system, a per-dimension GMM-based normalisation was used for all dimensions, including pitch.

Acoustic Segmentation/Clustering: For the acoustic segmentation and clustering an approach similar to the BN-E segmenter [4] was used. It consists of a GMM classifier to split the data into wideband speech, telephone speech, speech with music and pure music regions. The music is discarded and speech with music is treated as wideband speech. A Gender Dependent (GD) phone recogniser is then run to locate gender-change points and silence portions to enable these regions to be split into smaller segments. Two forms of clustering were then used. The first, as used in the BN-E system [4], was based on a GD top-down clustering scheme with arithmetic harmonic sphericity distance metric and occupancy based stopping criterion. This was used for the `v1` and `v2` segmentations. The second form of clustering used a symmetric divergence based change point detector and BIC agglomerative clustering which also refined the segmentation. The `v3` segmentation/clustering was based on this approach.

Seg	# Segments	# Clusters	Avg. Seg. Len.
v1	382	57	9.27 sec
v2	522	55	6.79 sec
v3	324	55	10.36 sec

Table 1. Number of segments and clusters on `eval04`.

Table 1 shows the number of segments and clusters produced for each of the three schemes for the `eval04` test set. Segmentation `v1` was generated using the standard settings from the BN-E segmenter. This was used for all the initial system development. The other segmentations, `v2` and `v3`, were tuned to produce more and fewer segments to give diversity in the segmentation.

Baseline Acoustic Models Performance: The baseline acoustic models, referred to as S1 in this work, were trained using the 147 hours of Mandarin only acoustic training data. Gender Independent (GI) decision tree clustered triphone HMMs with approximately 6K distinct states were estimated. The decision tree questions included tonal questions, allowing tonal triphone models to be generated. An average of 16 Gaussian components per state was used, the number assigned to each state was based on the state occupancy count. The standard form of parameter estimation, initially Maximum Likelihood (ML) followed by MPE [5] training, was used. Where GD models were generated the discriminative approach described in [6] was used. Unless otherwise stated GI MPE-trained acoustic models are used.

System		CER (%)		
		`dev04f`	`eval03m`	`eval04`
S1	HLDA	12.4	6.6	21.1
	GAUSS	11.9	6.2	20.3

Table 2. Unadapted performance of baseline systems using HLDA or GAUSS front-ends, V1 segmentation and lm1.0.

Table 2 shows the performance of the baseline systems using either the HLDA or GAUSS front-ends. In a similar fashion to the CTS systems, the use of Gaussianisation gave gains over HLDA. Over the test sets considered the gain was between 0.4% and 0.8% absolute. An interesting aspect of this baseline system is that no English[1] training data was used to generate the acoustic models (though English words were present in the dictionary and language model). If the English segments were ignored, the performance of the baseline system improved by 2.3% absolute on `dev04f` and by about 1% absolute on other test sets[2] .

Test Set	# Eng. Words	# Mand. Char.	Percentage
`dev04f`	147	8630	1.7
`eval03m`	92	8958	1.0
`eval04`	171	16163	1.0

Table 3. Amount of English words in different Mandarin test sets.

To investigate the level of English in the test data, the percentage of English for each test set was examined and shown in table 3 along withthe number of Mandarin characters in the reference transcription. The percentage of English is far larger than in the CTS task and this is clearly degrading the performance.

3. SYSTEM REFINEMENTS

Incorporating English Training Data: In the previous section the effect that the presence of the English data has on the performance was examined. One approach to handling this would be to automatically label segments of the data containing English and either not recognise that data, or use an English system for that segment. This is the approach adopted in [2]. However this approach was found to be unreliable. The scheme used in this work is to construct a system that recognises both English and Mandarin. As the vast majority of the data comes from Mandarin, the starting point was the baseline system from the previous section.

The first step was to expand the word-list to incorporate common English words. 5K English words were added to the baseline 50K word-list to give the 55K word-list. The Mandarin language was then rebuilt with the new word-list, this is referred to as `lm2.0`. The use of this updated word-list had no effect on recognition performance, as almost no English were produced in the hypothesis.

To improve the performance on the English data, a general English language model was constructed using the 55K word-list mentioned above. In addition, the 1 hour of English training data from the TDT4 Mandarin data as well as additional sub-sets from the TDT4 English data were added into the acoustic training data. The performance of these new systems using the HLDA front-end and

[1] When scoring Mandarin, hypothesised English words were deleted prior to scoring as these are meant to be marked as optional in the reference.

[2] This is the form of scoring that is sometimes presented, for example in [7].

English Data		Ratio Mandarin:English LM		
TDT4M	TDT4E	10:0	9:1	8:2
—	—	15.2	15.3	15.5
1hr	—	15.1	14.9	15.2
	10 hr	14.7	**13.9**	13.8
	50 hr	14.8	13.8	13.7
	100 hr	15.4	14.2	14.4

Table 4. %CER on `dev04f` using HLDA-ML models with different amounts of English acoustic data added from TDT4-Mandarin (TDT4M) and TDT4-English (TDT4E) sources and with the use of different interpolation weight ratios of the Mandarin-English LMs.

ML training only (for efficiency reasons) are shown in table 4. Simply interpolating the `lm2.0` language model with the general English language model gave no performance gain without adding English acoustic training data. Increasing the level of English acoustic training data upto about 50 hours gave gains in performance. The operating point selected for this work was to use the 1 hour of English data from the TDT4 Mandarin source, along with 10 hours of data from the TDT4 English data. This acoustic model will be referred to as the S2 acoustic model. For the language model an interpolation weight of 0.9 for the Mandarin LM and 0.1 for the English LM was used (`lm2.1`). It is interesting to note that the amount of English acoustic data used to get good performance gains was significantly greater than the percentage of English in any of the test sets.

Web Language Model Training Data: To increase the amount of language model training data, an additional 40M words of data were collected from the web. This data was split between about 34M words from broadcast sources (CCTV, NTDTV, VOA) and about 6M words from news paper sources. Separate language models were built for each of the two types of source and then interpolated with the language models described in section 2 to give `lm3.0`. In a similar fashion to the previous section, this Mandarin language was also interpolated with a general English language model with an interpolation weight ratio of 9:1 to form the `lm3.1` language model.

LM	Ratio M:E	Voc Size	PPlex (Inc/Ex Eng)		OOV(%) Inc/Ex Eng
			3-gram	4-gram	
lm1.0	-	50K	258/230	240/213	0.56/0.08
lm2.0	-	55K	270/230	250/213	0.17/0.08
lm2.1	9:1		259/246	240/227	
lm3.0	-	55K	190/165	178/154	0.17/0.08
lm3.1	9:1		188/178	176/166	

Table 5. Perplexities and OOVs on `dev04f` of different language models used in this work.

Table 5 shows the perplexities and Out Of Vocabulary (OOV) rates for the language models used in this paper on the `dev04f` test data. Two sets of numbers are quoted. The first includes English, the second excluding English. For all the LMs used the OOV rate excluding English was 0.08%. This could have been driven to zero by incorporating entries for every character in the lexicon, but this was not performed in this paper. The OOV rate including English dropped from 0.56% to 0.17% when the 55K word-list was used. Comparing the perplexities of using the web data or not, `lm3.1` against `lm2.1`, shows a significant drop in perplexity for both the tri-gram and 4-gram of about 60 to 70 points. Interpolating with the general English LM (for example `lm3.1` against `lm3.0`) shows a reduction in perplexity when English is included and, not surprisingly, a slight increase when English is excluded.

System	LM	CER (%)		
		`dev04f`	`eval03m`	`eval04`
S2 GAUSS	lm1.0	12.2	5.6	20.0
	lm2.0	12.3	5.2	20.2
	lm2.1	11.5	5.1	20.0
	lm3.1	9.7	5.0	18.7

Table 6. Unadapted Performance of MPE-trained S2 Models with a GAUSS front-end using different tri-gram LMs and v1 segmentation.

Given the large reduction of perplexity scores in table 5, it is interesting to evaluate the recognition performance with the language models. Table 6 shows the performance of the MPE-trained S2 models (including the 11 hours of English data) using the GAUSS front-end. The best performance was obtained with the `lm3.1` language model, yielding a 2.5% absolute reduction in CER on `dev04f` over the `lm1.0` language model.

Improved Acoustic Modelling: To further improve the acoustic modelling, and add additional possibility for combination in a multi-branch framework, two additional sets of acoustic models were built. The first used Speaker Adaptive Training (SAT), where separate speech and silence CMLLR transforms were used during training. The second was a SAT version of fMPE [8]. Here a global CMLLR transform was used during training, and the fMPE projection matrix estimated in this normalised space. For this fMPE system all adaptation, both CMLLR and MLLR in section 4 were run prior to the fMPE projection matrix being applied. This was found to yield slightly improved performance over other adaptation configurations.

4. SYSTEM COMBINATION

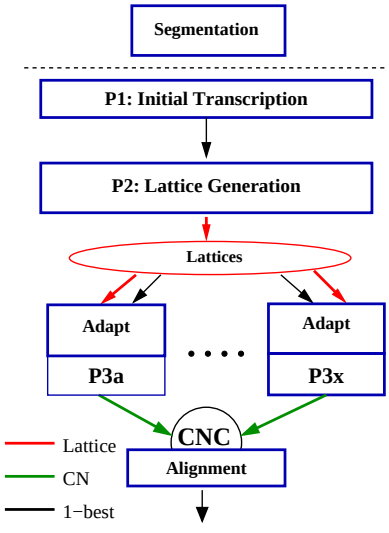

Fig. 1. BN-M multi-pass and multi-branch evaluation framework

In this section, the performance of various system combination

configurations was evaluated within a multi-pass/multi-branch framework. The basic structure of the system is shown in Figure 1. Initially, the audio data was automatically segmented and passed to the P1 stage where GI acoustic models were used to provide adaptation supervision for the P2 stage. Least squares linear regression and diagonal variance adaptation was then performed on GD models, which were used to generate lattices for subsequent rescoring. These lattices were generated with the `lm3.1` tri-gram language model and then the 4-gram language model applied. For the P3 stage, 1-best adaptation supervision and lattices from the P2 stage were used for CMLLR and lattice-based MLLR for both mean and full variance transforms. In the P3 stage all non-SAT models were GD. The final system output was derived by combining various P3 outputs using Confusion Network Combination (CNC). To add diversity to the system acoustic models using the compact 38 toneless phone set were also built. These also used 11 hours of English acoustic training data and are referred to as the S3 models.

Pass	System	CER (%)		
		`dev04f`	`eval03m`	`eval04`
P2-cn	S2 HLDA	8.4	4.7	17.6
P3b-cn	S2 HLDA	7.9	4.4	17.0
P3d-cn	S2 GAUSS	7.8	4.3	16.6
P3e-cn	+SAT	7.5	4.0	16.4
P3f-cn	+fMPE	7.1	4.1	16.1
P3g-cn	S3 HLDA	7.7	4.4	16.9
P3h-cn	S3 GAUSS	7.3	4.4	16.5
P3i-cn	+SAT	7.6	4.4	16.5
P3d+P3h	CNC	7.3	4.0	16.3
P3f+P3h		7.1	4.0	16.2

Table 7. CER in P2, various P3 branches and confusion network combination in the development framework using v1 segmentation and `lm3.1`.

The performance of the individual acoustic models and their combinations is summarised in Table 7. For both S2 and S3 systems, Gaussianisation again gave consistent improvements over the baseline HLDA front-end. For the S2 system, SAT yielded a further 0.2–0.3% absolute reduction. The best combination gains were obtained by combining branches with different phone sets. For example, the combination of the S2 and S3 GAUSS systems gave a 0.3% and 0.2% absolute improvements on `eval03m` and `eval04` respectively. However, combining the best single branch performance, the S2 GAUSS+SAT+fMPE system, with any other branch gave no gains. This may be due to the large performance gap between the S2 GAUSS+SAT+fMPE system and the other systems.

Segmentation	CER (%)		
	`dev04f`	`eval03m`	`eval04`
v1 (P3f-cn)	7.1	4.1	16.1
v2	7.0	4.5	16.1
v3	7.3	3.9	16.0
v1 $\oplus$ v2 (ROVER)	7.0	4.1	16.0
v1 $\oplus$ v3 (ROVER)	6.9	3.9	15.9

Table 8. CER of dual segmentation system. The S2 system uses GD MPE models in P2 and GAUSS+SAT+fMPE models in P3.

Finally, different acoustic segmentations were also investigated as this was found to yield gains for BN-English [6]. Single P3 branch branches using the S2 GAUSS+SAT+fMPE models set with each of the three segmentations described in section 2 were evaluated. These results are shown in table 8. By combining two segmentations using ROVER [9] slight gains in performance were obtained. The best dual segmentation system, which combined the v1 and v3 segmentations, reduced the CER by 0.1% to 0.2% absolute. The best performance on the `eval04` test set was 15.9%, which shows state-of-the-art performance when compared with previously published results on this task [2, 7].

5. CONCLUSIONS

This paper has described the development of a Mandarin broadcast news transcription system. An effective approach to address the issue of decoding English speech present in primarily Mandarin BN data was presented by incorporating English acoustic model data and language model data in training. The transcription system uses system combination strategies and the benefits of including alternate phone sets and acoustic model types has been discussed. A small further benefit was available from the use of multiple segmentations. The overall system delivers state-of-the-art performance on all the test sets considered.

6. REFERENCES

[1] M. J. F. Gales, B. Jia, X. Liu, K. C. Sim, P. C. Woodland, and K. Yu, "Development of the CUHTK 2004 Mandarin Conversational Telephone speech transcription systems," in *Proc. ICASSP*, Philadelphia, PA, March 2005.

[2] H. Yu, Y. C. Tam, T. Schaaf, S. Stuker, Q. Jin, M. Noamany, and T. Shultz, "The ISL RT-04 Mandarin Broadcast News evaluation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, November 2004.

[3] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, Montreal, Canada, March 2004.

[4] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang, and P. C. Woodland, "Recent advances in broadcast news transcription," in *Proc. ASRU Workshop*, November 2003, pp. 105–110.

[5] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, Orlando, FL, May 2002.

[6] D. Y. Kim, H. Y. Chan, G. Evermann, M. J. F. Gales, D. Mrva, K. C. Sim, and P. C. Woodland, "Development of the CU-HTK 2004 Broadcast News transcription systems," in *Proc. ICASSP*, Philadelphia, PA, March 2005.

[7] M. Afify, L.Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic Broadcast News transcription at BBN," in *Proc. Eurospeech*, Lisbon, Portugal, September 2005.

[8] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, 2005.

[9] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. ASRU Workshop*, 1997.