Improvements to the BBN RT04 Mandarin Conversational Telephone Speech Recognition System

Jeff Z. Ma, Spyros Matsoukas

BBN Technologies, Cambridge, MA, USA
{jma,smatsouk}@bbn.com

Abstract

BBN's 20 times real-time (20xRT) Mandarin conversational telephone speech (CTS) recognition system achieved the lowest character error rate (CER) in the Rich Transcription 2004 fall (RT04F) evaluation conducted by NIST. This paper focuses on the work we have done after the evaluation. The work includes porting of more new acoustic modeling technologies we had developed on English, such as long-span features, a modified HLDA-SAT, etc., diagnoses of the problems we had encountered in the evaluation, such as problems in pitch, silence chopping and automatic segmentation, and solutions we found for the problems. With all these new technologies and problem solutions incorporated and a new design of the 20xRT system architecture we achieved a 2.1% absolute reduction in CER on the RT04 evaluation test set.

1. Introduction

The RT04F evaluation was conducted by NIST for the EARS program in the fall of 2004 [1]. It evaluated speech recognition systems on CTS and broadcast news (BN) domains for three languages: English, Mandarin and Arabic. It required all the systems to run no slower than 20xRT for CTS and 10xRT for BN. BBN participated in both the CTS and BN domains for all the three languages. During the development period we focused primarily on the English systems and spent a short period of time before the evaluation deadline to port the new technologies from English to the other two languages and build systems for them.

The new technologies, that were verified to work well on English, include long-span features obtained by frame concatenation, a new HLDA-SAT training procedure on the frame-concatenated features, and a new procedure generating lattices for the discriminative training directly from the backward decoding pass. We were only able to use the long-span features in one of the Mandarin evaluation models (the MPE-HLDA model).

During the development of the Mandarin system we encountered several problems. The first problem was poor cepstral normalization caused by long silence periods in the new HKUST data. The second one was the abnormal behavior of the pitch features. The third one was an out-of-memory problem during the lattice annotation for discriminative training. To save time, we simply skipped them and switched to sub-optimal replacements for them. After the evaluation we have done more work on both the porting of the new technologies and the investigation of the problems. This paper reports such work.

The paper is organized as follows: a summary of the RT04 Mandarin CTS system is given in Section 2; Section 3 mainly focuses on the porting of new technologies; Section 4 reports the investigation on pitch features; Section 5 addresses our improvement on the automatic segmentation algorithm; we report our further work on silence chopping in Section 6; finally, we present a new updated 20xRT system in Section 7.

2. RT04 Mandarin CTS System Summary

The BBN RT04 20xRT Mandarin CTS system used 85 hours of acoustic training data consisting of a new 50-hour corpus collected and transcribed by Hong Kong University of Science and Technology (HKUST), and the old 35-hour Callhome and Callfriend data. Language models were trained on a total of 141 millions of words (about 1 million words from the acoustic transcripts, 117 million Web conversational words provided by the University of Washington, and 123 million words from broadcast news).

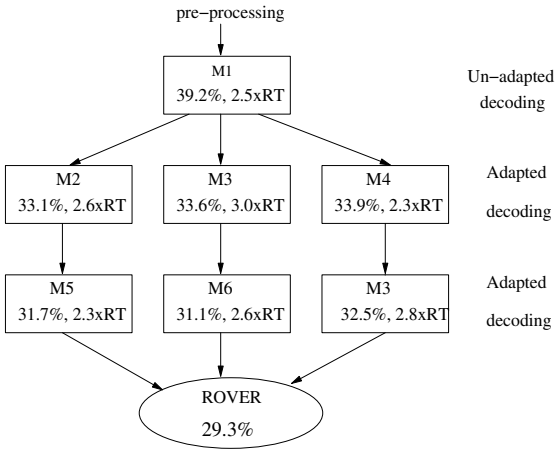

Figure 1: BBN's RT04 20xRT Mandarin CTS system architecture (CERs and speeds measured on Eval04)

The system architecture is shown in Figure 1. We trained 6 different gender-independent (GI) models for the cross-model adaptation and system combination. All the models were estimated so as to minimize expected phoneme error (MPE) on the training data, with the exception of model M1, which was trained with Maximum Mutual Information (MMI). All the MPE models were speaker adaptively trained using HLDA-SAT [3]. Model M1 was not speaker adaptively trained because it was used in the un-adapted decoding. The model denoted as "M6" used MPE-HLDA [2] to project the concatenated frames, and all other models used HLDA to project the basic frame and its 1st, 2nd, and 3rd derivatives. In all cases, the final feature space consisted of 46 dimensions. The six models permuted mainly in three aspects: feature (PLP vs. MFCC), use of mixture exponents (yes vs. no) and phoneme set (77 phonemes vs. 147 phonemes). We found that

those three variations as well as the use of MPE-HLDA benefited the cross-model adaptation and combination greatly.

3. Porting New Technologies

3.1. A new phoneme set

During the Mandarin evaluation, we observed that our lattice-based discriminative training procedure ran out of memory for hundreds of utterances during the crossword annotation of lattices that were generated directly from the backward decoding pass. We found that the problem was caused by the large number of single-phoneme words in Mandarin. Lattices containing many paths with sequences of single phoneme words increased in size dramatically after the crossword quinphone expansion, thereby requesting excessive amounts of memory during the Viterbi annotation. Unfortunately, we didn't have time to pin down this problem. We simply switched back to the old procedure that generated lattices from n-best lists (300-best).

There are hundreds of Mandarin characters uttered as single phonemes. However, each of them can be represented with one of five vowel phonemes, "a", "i", "o", "e", and "er", with different tones. We inserted five dummy phonemes, "Da", "Di", "Do", "De" and "Der", in front of the five phonemes, respectively, to turn them into dummy double phonemes [5]. In this way, all the single phoneme words become double-phoneme words. For example, a word with the pronunciation "a" in the original dictionary has the pronunciation "Da-a" in the new dictionary. In this way the total number of crossword quinphones was reduced by 40%, which enabled our new lattice generation procedure to run with reasonable memory requirements. Adding the five dummy phonemes to our original 77-phoneme set, we ended up with an 82-phoneme set.

Table 1: Effect of the new phoneme set on CER (measured on the Dev04 test set)

Phoneme set	Un-adapted CER	Adapted CER
77	40.5%	36.1%
82	40.4%	36.0%

A comparison of the two phoneme sets listed in Table 1 shows that the new 82-phoneme set performs as well as the old 77-phoneme set. All our experiments afterward used the new phoneme set.

3.2. Frame concatenation

We had found on English that the use of long-span features, obtained by concatenating consecutive frames, outperformed our regular derivative features that include the basic frame and its 1st, 2nd, and 3rd derivatives [2]. The derivative calculation can be viewed as a pre-determined linear feature transformation over a span of 9 frames. In contrast, in the concatenation case, a general linear transform is estimated from the concatenated frames by using LDA. The concatenation can easily span more then 9 frames. We carried out similar experiments on Mandarin.

Table 2 lists those experiments, where the "basic frame" includes the normalized energy, the first 14 PLP cepstrals, pitch (F0), and probability of voicing (PV). For both the derivative features and the concatenated ones, LDA+MLLT was used to project them down to 46-dimensional features. It shows that the 9-frame concatenation outperforms the derivative features by 0.5% absolute. Concatenating more frames (the last row of Table 2) doesn't yield further gain. Therefore, we chose the 9-frame concatenation for our later experiments.

Table 2: Frame concatenation experiments on 85-hour Mandarin data (CER measured on Dev04 set)

LDA input feature	Un-adapted CER
basic frame + derivs.	40.9%
concat. 9 basic frames	40.4%
concat. 13 basic frames	40.6%

3.3. A modified version of HLDA-SAT

We had developed the HLDA-SAT model in the RT03 evaluation when we were using the derivative features. In that implementation, a speaker-dependent constrained MLLR (CMLLR) transform was first applied to the derivative features, and then a global HLDA projection was estimated in the transformed space. The global HLDA was then combined with a second set of speaker dependent CMLLR transforms, producing speaker dependent feature projections. Finally, a regular SAT model was estimated in the resulting space. We found that the HLDA-SAT gave 0.7% absolute gain over regular SAT on the RT03 Mandarin training data (35 hours). We observed a similar gain with the RT04 Mandarin training set (85 hours). Considering the frame-concatenation gain obtained on the SI models, we certainly wanted to train the HLDA-SAT model on the frame-concatenated features as well. However, the high dimensionality of the concatenated frames makes the first step of the HLDA-SAT model training susceptible to over-fitting. In order to take advantage of the long-span features from concatenation as well as avoid the high dimensionality problem, we modified the first step of HLDA-SAT for the frame-concatenated case. Rather than estimating the speaker dependent CMLLR transforms on the concatenated frames, we estimated the adaptation transforms on the basic frames. These transforms were applied to the basic frame features prior to frame concatenation for the global HLDA projection. Experiments done on the English data showed that this modified HLDA-SAT training procedure retains most of the frame-concatenation gain observed on SI models.

Table 3 shows the results of similar experiments we did on the Mandarin data. Comparing the second and third rows, we can see that the modified HLDA-SAT on the frame-concatenated features gives a 0.8% absolute gain, which is larger than the absolute improvement observed on the SI model (0.5%). This means that the modified HLDA-SAT model works well on the frame-concatenated features on Mandarin. Both experiments in the second and third rows used the state-clustering from their SI model training, even though they were being trained on the transformed spaces.

The experiment given in the last row did the same thing as the one in the third row, but it re-did the state clustering after the basic frames were transformed. It shows that the state re-clustering provides 0.4% absolute gain. A similar re-clustering on the old HLDA-SAT case didn't affect the CER.

Table 3: Performance (CER measured on Dev04) of the modified HLDA-SAT model

LDA Input feature	HLDA-SAT	Adapted CER
basic frame + derivs.	old	36.8%
concat. 9 basic frames	modified	36.0%
concat. 9 basic frames	modified (state re-clust.)	35.6%

3.4. MPE training on top of the frame-concatenated HLDA-SAT model

With the successful combination of the frame concatenation and the HLDA-SAT model training, we continued the MPE training on the HLDA-SAT model. Recall that we have been using the new phoneme set in our experiments, which greatly alleviates the lattice crossword expansion problem we encountered during the Mandarin evaluation. So, this time we used the new discriminative training procedure that employs lattices directly generated out of the decoder's backward pass.

Table 4: Performance (CER on Dev04 test set) of the MPE model trained on frame-concatenated frames

Feature Set	MPE training	Adapted CER
basic frame + derivs.	Old procedure	34.8%
concat. 9 basic frames	new procedure	33.3%

The results are shown in Table 4, where the derivative MPE model (the second row) was trained on top of the derivative HLDA-SAT model, and the frame-concatenated MPE (the third row) on top of the frame-concatenated HLDA-SAT with state re-clustering. Compared with their HLDA-SAT models the MPE models provide 2.3% CER reduction in the frame concatenation case and 2.0% in the derivative case. So, the gain from frame concatenation has been well preserved after MPE training. The 0.3% extra MPE gain in the frame-concatenated case is probably due to the use of the new MPE training procedure (mainly the use of deeper lattices).

The derivative MPE model in Table 4 is one of the models used in the RT04 system. So, we obtained 1.5% absolute gain after porting the new technologies.

4. Pitch Features

Pitch features are of importance for tonal languages such as Mandarin. We have included the pitch and PV features in the basic frames. For the RT04 Mandarin evaluation we tried a new pitch algorithm that is similar to IBM's algorithm [4] and found that the new algorithm smoothes the un-voicing parts better than our old one. An initial experiment also showed that the new pitch outperformed the old one in terms of CER. Hence, we used the new algorithm in the RT04 evaluation.

After the evaluation we did a thorough comparison of the two algorithms, shown in Table 5. We can see that on the derivative features (the 2nd and 3rd rows) the old pitch outperforms the new one by 0.6% after adaptation. We also compared them for the frame concatenation case (the fourth and fifth rows), and found that they made no difference after adaptation. Overall, the old pitch algorithm gave better (or equal) CER reduction than the new one. Besides different smoothing, another difference is that the new pitch values are in the log domain. So smoothing un-voiced parts better and working in log domain doesn't help decrease CER.

Table 5: A comparison of the old and new pitch algorithms (CER measured on Dev04)

Pitch	Feature	Adapted CER
Old	Basic frame + derivs.	36.2%
New	Basic frame + derivs.	36.8%
Old	Concat. 9 basic frames	35.6%
New	Concat. 9 basic frames	35.6%

5. Automatic Segmentation

We observed that our RT04 Mandarin system performed much better (>1% absolute) on Dev04, the development test set, and only marginally better (0.3%) on Eval04, the evaluation test set, compared to other participants in the RT04 evaluation. It prompted us to investigate the reasons after the evaluation. We found that our automatic segmentation algorithm caused a much larger degradation on the evaluation test set. The effect of automatic segmentation on the evaluation test set was measured against a manual segmentation that we generated from the Eval04 STM (segment time marked) file released by NIST. Table 6 shows the degradation from automatic segmentation in un-adapted decoding on the Dev04 and Eval04 test sets (the third column). We lost 2.2% absolute in CER due to the automatic segmentation on Eval04, which is significantly larger than the 0.6% loss on Dev04.

Table 6: Degradation caused by the automatic segmentation on Dev04 and Eval04 test sets

Test set	Manu. seg.	RT04 Auto. seg. (loss)	New auto. seg. (loss)
Dev04	38.6 %	38.0% (0.6%)	38.0% (0%)
Eval04	39.2 %	37.0 % (2.2%)	37.6% (0.6%)

Closer comparison of the manual and automatic segmentation errors revealed that on Eval04 the automatic segmentation caused a large number of insertion errors. The insertion error (3.4%) in recognition with automatic segmentation is much higher than that with manual segmentation (1.9%). To understand the reasons, we listened

to a subset of the conversations in the Eval04 set, where the automatic segmentation had resulted in high insertion errors. We found that most of these conversations have severe cross-talk, and the rest strong background noises. Apparently our segmentation algorithm treated cross-talk as regular speech on both sides of the conversation, which resulted in an increased number of insertions.

For CTS automatic segmentation, we used a 4-state ergodic HMM with Gaussian mixture models (GMMs) for the state distributions [6] to model the 4 scenarios in the two-channel conversations: speech on both channels (SS), noise on both channels (NN), speech on one channel and noise on the other one (SN and NS). Diagnosis revealed that cross-talk data, which should belong to either NS or SN, was mis-assigned to SS during the GMM training. It caused the cross-talk appearing in the test data to be recognized as SS, thereby causing insertion errors. After this discovery we tried a simple straightforward method to correct the labeling of the training data. We did one extra processing on the SS data, and re-assigned the data to either SN or NS class if the channel correlation was higher than a threshold (0.27 was the best we found). The performance of this modified automatic segmentation algorithm is given in the last column of Table 6. We can see that the extra processing on the SS data reduced the degradation significantly on Eval04 (from 2.2% to 0.6%). So, the modified algorithm is efficient to recover the loss due to cross-talk.

6. Silence Chopping

We found that the HKUST training data has long silences. We chopped the extra silences at utterance endpoints, based on the forced alignments, during the evaluation. It gave us 2% absolute gain. After the evaluation we further investigated this issue. We set up an automatic procedure to chop long silences. Experiments show that this new automatic chopping procedure yields 0.2-0.3% absolute gain over the old manual chopping.

7. An Updated System

We retrained some of the models used in the RT04 evaluation using the improvements described previously. as well as one new model. We also found that it worked slightly better when we moved one model from the 1st adapted decoding pass to the 2nd pass in Figure 1. So we have modified the RT04 system architecture slightly. The new 20xRT system architecture is demonstrated in Figure 2. Seven different models are used in the new system, and they are all MPE models. The model denoted as "M1r", "M4r" and "M5r' are the retrained versions of "M1", "M4" and "M5" in Figure 1, respectively, with all the improvements incorporated. The model "M7r" is a new model trained similarly to "M5r" but with the 147-phoneme set. The models "M2", "M3" and "M6", are the same ones as in Figure 1. Recall that these three models were trained with the "new" pitch features. We found that keeping those models unchanged benefits the cross-model adaptation and system combination, since they used different pitch features.

This new system turned out 27.2% CER after the final system combination, which is 2.1% absolute lower than the RT04 system.

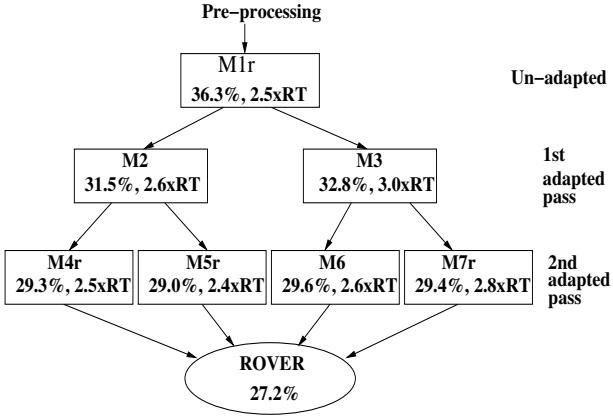

Figure 2: An updated 20xRT system architecture (CERs and speeds measured on Eval04)

8. Conclusion

In this paper we have described the work we had done on CTS Mandarin after the EARS RT04 evaluation. The work includes the porting of new acoustic modeling technologies from English, the diagnoses of the problems we came across during the evaluation and the solutions for them. We have ported the new technologies of using long-span features on all the SI, HLDA-SAT and MPE model training. These technologies yielded 1.5% absolute CER reductions, so they work successfully on Mandarin. The thorough comparison of two pitch algorithms showed that the better smoothing of un-voicing parts and working in log domain doesn't help CER reduction. We have analyzed the loss of automatic segmentation on Eval04 and worked out a method to improve the algorithm. We have also further looked into the silence chopping issue, which had not been fully investigated in the evaluation. We verified that the new automatic chopping procedure we set up after the evaluation worked better. Finally, we have designed a new 20xRT system that incorporated all the improvements from both the new technologies and the improved algorithms. The system achieved 2.1% absolute CER reduction after final system combination.

9. References

[1] http://www.nist.gov/speech/tests/rt/rt2004/fall

[2] B. Zhang et al., "Long Span Features and MPE HLDA", Rich Transcription Workshop, Palisades, NY, 2004.

[3] S. Matsoukas and R. Schwartz, "Improved Speaker Adaptation Using Speaker Dependent Feature Projections," *ASRU 2003*.

[4] C. J. Chen et al., "New Methods in Continuous Mandarin Speech Recognition", *Eurospeech 1997*, pp1543-1546, Rhodes, Greece, 1997.

[5] J. Zhang et al., "Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition", Eurospeech, pp1617-1620, 2001, Aalborg, Denmark.

[6] D. Liu and F. Kubala, "A Cross-Channel Modeling Approach for Automatic Segmentation of Conversational Telephone Speech", ICASSP 2003.