PROGRESS IN THE BBN 2007 MANDARIN SPEECH TO TEXT SYSTEM

Tim Ng, Bing Zhang, Kham Nguyen† and Long Nguyen

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
† Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
{tng,bzhang,knguyen,ln}@bbn.com

ABSTRACT

In this paper, we describe the BBN 2007 Mandarin Speech-to-Text system developed for the GALE Evaluation 2007. In comparison to the BBN 2006 Mandarin system, we achieved 25% relative reduction in character error rate on the most important test sets. The utilization of all available training data provided the largest contribution to the improvement. The use of a better pitch tracking algorithm also contributed significantly, while system combination made some noticeable improvement too.

Index Terms— Speech recognition, Mandarin, pitch, system combination

1. INTRODUCTION

This paper presents the BBN 2007 Mandarin Speech-to-Text (STT) system (BBN07) which was used to transcribe the Mandarin audio data into Chinese text for the GALE Evaluation 2007. The audio data is categorized into Broadcast News (BN) and Broadcast Conversation (BC) to reflect the different degree of difficulty, and the classification is supposed to be hidden during decoding. The majority of the speech in BN is spoken in a formal or professional news reporting style, while BC consists of interviews, discussions, and talk-shows.

The development of BBN07 started with the inclusion of the 850 hours of acoustic data and 200 million characters of transcripts that were released after the GALE Evaluation 2006. The expansion of the training set provided 17% relative reduction in Character Error Rate (CER). To exploit the linguistic information for Mandarin [1], we adopted an improved pitch extraction algorithm. The use of the Robust Algorithm for Pitch Tracking (RAPT) [2] pitch followed by a smoothing and normalization procedure further improved the system by 8% relative in CER. In addition to the primary system, two sets of Region Dependent Transform (RDT) [3] acoustic models (AM) were constructed for system combination purposes. To provide a compromise for BN and BC data, two sets of audio segmentations were derived to reflect the difference in utterance length for BN and BC. Therefore, BBN07 consists of 6 sub-systems and the resulting hypotheses are combined at the end using ROVER [4]. The system combination strategy contributed about 3% relative improvement. We achieved 25% relative improvement in total.

This work was supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022

The paper is organized as follows: Section 2 summarizes the acoustic and language training corpora and the test data. The Mandarin STT system, including the construction of the acoustic and language models, are described in Section 3. Section 4 comprises the presentation and discussion of experiment results, while Section 5 presents the conclusion.

2. TRAINING CORPORA AND TEST DATA

2.1. Acoustic Training Data

As shown in Table 1, the BBN07 was trained on a 1371-hour acoustic training set that consists of: 127-hour acoustic data from the LDC Hub-4 Mandarin (Hub4M) and TDT4 corpora, 135-hour data from TDT2 and TDT3, 663-hour data from GALE Phase 1, and 446-hour data from GALE Phase 2. As various errors were found in LDC's Quick Transcriptions, data from most of these corpora was automatically selected using light supervision [5]. Data from the Hub4M, TDT2, TDT3 and TDT4 is considered BN, while approximately half of the data in GALE Phase 1 and Phase 2 is BC. The BBN 2006 Mandarin system (BBN06) was trained on a 520-hour subset as GALE Phase 2 and the second part of the Phase 1 were not available then.

Subset	BN	BC	Total
Hub4M, TDT4	127	–	127
TDT2, TDT3	135	–	135
GALE Phase 1	333	330	663
GALE Phase 2	200	246	446
Total	795	576	1371

Table 1. Acoustic training data in hours

2.2. Language Training Data

The language model (LM) training corpus for BBN07 consists of 3.5 billion characters (see Table 2). A small portion

of the LM data comes from transcripts of audio data, while the majority is web data collected from the Internet by different sites including BBN, Cambridge University (CU) and the University of Washington (UW). These web data were collected from various sources, such as China National Radio (CNR), China Central Television (CCTV), Voice of America (VOA), People's Daily, Xinhua News, ZaoBao, Phoenix TV, British Broadcasting Corporation (BBC) and New Tang Dynasty Television (NTDTV). These LM texts span over a long period, from 1991 to 2007. Data from certain months (Dec 2001, Feb 2001, Nov 2003 and Apr 2004) was excluded to prevent overlapping with the date of the development and evaluation test sets. The LMs used in BBN06 was trained on a 3.3-billion-character subset because the transcripts from the GALE Phase 2 and the second part of the Phase 1 were not released.

Source	Epoch	# Characters
Hub4M	1997	1.9M
TDT2, TDT3, TDT4	1998, 2000-2001	27.0M
GALE Phase 1,2	2004-2007	36.4M
LDC Gigaword2	1991-2004	1217.9M
BBN web data	1998-2005	1437.2M
CU web data	2000-2006	495.4M
UW web data	2001-2005	272.2M
Overall	1991-2007	3.5B

Table 2. Breakdown of the language training data

2.3. Test Data

Five test sets were used throughout the system development. The development set designed by LDC for Mandarin STT development for the 2007 GALE Evaluation, Dev07, includes 74 stories (32 CCTV, 28 Phoenix and 14 NTDTV) aired in Nov 2006 with a total duration of 2.4 hours. The 2006 GALE Evaluation test set is denoted as Eval06. The BC tuning set, bcmt05, consists of 1.5 hours of BC data collected in Mar 2005. The BN tuning set, bnmt06, consists of 1.8 hours of BN data, collected in Feb 2001, Nov 2003, Apr 2004 and Oct 2005. ct6, the union of bcmt05 and bnmt06, serves as a tuning set for parameter optimization in the system development.

The 2007 GALE Evaluation test data, denoted as Eval07,was unseen during the system development. Eval07 consists of 83 TV programs (2 from ANHUI TV, 50 from CCTV, 11 from NTDTV and 20 from Phoenix TV). They were all aired in Nov and Dec 2006 with a total duration of 2.4 hours.

3. MANDARIN SPEECH-TO-TEXT SYSTEM

The recognition process in our system is divided into three stages: audio segmentation, feature extraction, and decoding.

3.1. Audio Segmentation

Speech data is automatically segmented and grouped into speaker clusters using the same process as described in [6]. The audio segmentation which produces utterances of an 8-second length on average was tuned mainly for the BN. Such audio segmentation is used in BBN06 and the primary system of BBN07.

3.2. Feature Extraction

Acoustic feature extraction is performed on the segments produced in audio segmentation. The length of each speech frame is 25 ms with a frame rate of 100 frames/sec. For each frame, 14 perceptual linear prediction (PLP) [7] derived cepstral coefficients, energy and log pitch are extracted. The pitch tracking algorithm as described in [6] was used in BBN06 while the RAPT algorithm is incorporated in BBN07. Rather than using derivatives, the Long Span Features (LSF) [8] are employed. The 9 successive frames of steady features (centered at the current frame) are concatenated. This block of features is projected onto a 60-dimensional feature space using Linear Discriminant Analysis (LDA).

3.3. Decoding

In general, the BBN decoding stage comprises three decoding passes. The first pass, unadapted decoding (UDEC), provides supervision for model adaptation and decoding in the second pass (ADEC-0), which in turn provides supervision for the third pass (ADEC-1). The first pass uses a speaker independent (SI) model, while the subsequent two passes use a speaker adaptive training (SAT) model. The standard BBN decoding strategy was described in [6] but the speaker adaptation process has been changed due to the use of LSF.

The Heteroscedastic Linear Discriminant Analysis-Speaker Adaptative Training (HLDA-SAT) transformation described in [9] is used in adaptation. A pre-transform is first estimated on the steady features for each speaker cluster using constrained maximum likelihood linear regression (CMLLR) [10]. LSF features are then obtained based on the pre-transformed features and LDA projections. Finally, a post-transform is computed on the LSF features using CMLLR again.

3.4. Acoustic Modeling

The BBN Mandarin STT system uses phonetic hidden Markov models (HMMs) to represent each of the 76 phonemes [6]. Each HMM has a left-to-right topology and consists of 5 states. These phonetic models are used as building blocks for words from the decoding lexicon. A two-level tying structure is used for the means and variances, and the mixture weights.

In a conventional BBN decoding system, a total of six AMs are needed for the decoding passes. This includes the State Tied Mixture (STM), the state-clustered noncross-word quinphone model (SCTM-NX) and the state-clustered cross-word quinphone model (SCTM-X) for both SI and SAT. Although the final models are discriminatively trained using Minimum Phone Frame Error (MPFE) [11], maximum likelihood (ML) models are also trained as initial models to build lattices required for MPFE training, and to obtain mixture weights. In BBN07, the SCTM-X is the largest model and it consists of 1.3 million Gaussians.

3.5. Language Modeling

Language models were built for a 65K decoding lexicon which was used in both BBN06 and BBN07 constructed by expanding the 48K dictionary used in [6] based on the occupancy frequency in the 3.3 billion-character language training corpus. Word segmentation was performed on all data in the language training corpus using the Longest Substring Match algorithm based on the 65K decoding lexicon. The training corpus was then partitioned into 22 groups according to their genre and sources. An LM was trained on each of the 22 groups using the modified Kneser-Ney smoothing. The 22 LMs were then linearly combined with weights optimized on ct6 using the EM algorithm. The bi- and tri-gram LMs were pruned to be tractable for the forward and backward decoding passes while an unpruned 4-gram LM is used in lattice rescoring. In total, there are 43M bi-grams, 100M tri-grams and 888M 4-grams in our language models.

4. EXPERIMENT RESULTS

This section reports the improvement in CER by using all available training data, the improved pitch features, and the complementary systems for combination.

4.1. More Training Data

As shown in the first and second row of Table 3, we obtained 17% relative reduction in CER for Dev07 (13.9% vs. 11.6%) by expanding the AM training set from 520 to 1371 hours and LM from 3.3 to 3.5 billion characters through adding the new data from GALE Phase 1 and 2. Our detailed analysis showed that the improvement comes almost entirely from the expansion of the AM training set.

4.2. Improved Pitch Feature

The pitch extraction algorithm used in [6] and BBN06 does not handle halving and doubling errors, and generates random pitch values for unvoiced regions. The RAPT algorithm employs dynamic programming to clean up the errors, and unvoiced regions are also detected in the process. We incorporated the RAPT algorithm in BBN07 for voiced/unvoiced detection and pitch extraction for the voiced regions. The log pitches for the unvoiced regions are linearly interpolated. A 75-frame moving window mean normalization is then applied to compensate speaker and phrase effect [12].

The third row in Table 3 represents the system in which the RAPT pitch features are used while the scond row corresponds to the old pitch system. The use of the RAPT pitch features provided 8% relative reduction in CER for Dev07 (11.6% vs. 10.7%).

System	bnmt06	bcmt05	Eval06	Dev07
BBN06	8.8	18.3	18.9	13.9
+ new training data	8.0	16.8	17.2	11.6
+ RAPT Pitch	7.7	16.3	16.5	10.7

Table 3. CERs for the systems using more training data and different pitch features

4.3. System Combination

We developed two complementary systems in addition to the primary MPFE system: RDT, and its variation direct-RDT or dRDT. RDT is a feature transformation method in which a discriminative training criterion (e.g. MPFE) is used to optimize a set of linear projections, with each region-dependent projection concentrated on a cluster of HMM states. The RDT is first trained together with an SCTM cross-word model (both the transformation and the AM are updated in the training), then an SAT MPFE system is trained on the features after the RDT transformation. In the dRDT training, only the feature transform is optimized for the RDT cross-word MPFE model. More technical details can be found in [3].

The primary audio segmentation of 8-second average length was mainly tuned for BN data, and another set of 4-second segmentations was derived to reflect the fact that the average utterance length is shorter for BC data. These two sets of segmentation are used in parallel to provide a compromise for both BN and BC data.

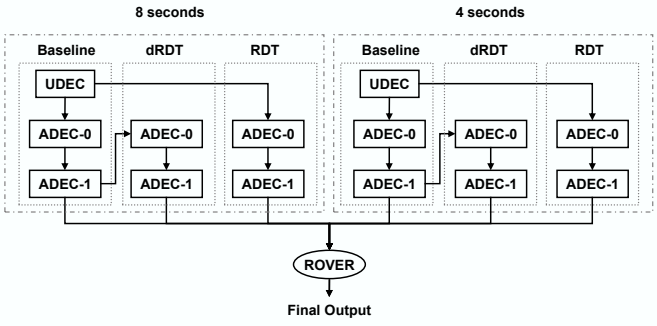

Fig. 1. Evaluation 2007 BBN Mandarin STT schematic diagram

As shown in Figure 1, BBN07 is the combination of the primary MPFE and two complementary RDT systems with the dual audio segmentations. The hypotheses from the SI decoding of the MPFE system were fed to the RDT system while the final hypotheses of the MPFE system were fed to the dRDT system. The input audio data is decoded 6 times with the 3 sets of AMs and 2 sets of audio segmentations, and the resulting hypotheses are then combined at the end using ROVER.

The results given in Table 4 show that the performances for the primary MPFE and its two complementary systems are competitive. The results for the 6 sub-systems and the combined system (shown in Figure 1) are given in Table 5. These results show that a 3% to 5% relative reduction in CER for the test sets was achieved using the system combination as compared to the primary system. In comparison to BBN06, BBN07 achieved about 25% relative improvement for Dev07 and Eval07 (see Table 6).

System	bnmt06	bcmt05	Eval06	Dev07	Eval07
MPFE	7.7	16.3	16.5	10.7	10.6
RDT	7.6	15.9	16.2	10.6	10.3
dRDT	7.6	15.8	16.0	10.7	10.4

Table 4. CERs for the MPFE, RDT and dRDT decoding systems

System	Eval06	Dev07	Eval07
MPFE, 8 seconds (primary)	16.5	10.7	10.6
RDT, 8 seconds	16.2	10.6	10.3
dRDT, 8 seconds	16.0	10.7	10.4
MPFE, 4 seconds	16.5	11.0	10.6
RDT, 4 seconds	16.5	10.8	10.5
dRDT, 4 seconds	16.4	10.9	10.5
BBN07 (ROVER)	15.8	10.4	10.1

Table 5. CERs for the 6 sub-systems and the combined system

System	Eval06	Dev07	Eval07
BBN06	18.9	13.9	13.4
BBN07	15.8	10.4	10.1
Relative Improvement	16.4	25.2	24.6

Table 6. Improvement in CERs for BBN07 and BBN06

5. CONCLUSION

We have presented the BBN 2007 Mandarin Speech-to-Text system for the GALE Evaluation 2007. In comparison to the BBN 2006 Mandarin STT system, the BBN 2007 Mandarin STT system achieved 25% relative reduction in CER for most representative test sets. The utilization of all available data provided a 17% relative gain. An 8% relative improvement in CER was obtained through the use of an improved pitch tracking algorithm followed by a procedure of smoothing and normalization. It is also demonstrated that a further 3% to 5% relative improvement can be obtained through system combination.

6. REFERENCES

[1] T. Ng, M. Siu, and M. Ostendorf, "A quantitative assessment of the importance of tone in Mandarin speech recognition," *Signal Processing Letters*, vol. 12, no. 12, pp. 867–870, 2005.

[2] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, pp. 495–518, 1995.

[3] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region-dependent transform for speech recognition," *ICASSP*, 2006.

[4] J. G. Fiscus, "A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER)," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–354, 1997.

[5] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *ICASSP*, May 2004.

[6] B. Xiang and L. Nguyen et al, "The BBN Mandarin broadcast news transcription system," *InterSpeech*, pp. 1649–1562, September 2005.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87(4), pp. 1738–1752, 1990.

[8] B. Zhang, S. Matsoukas, and R. Schwartz, "Long span features and minimum phoneme error heteroscedastic linear discriminant analysis," *Proceedings of EARS RT-04 Workshop*, 2004.

[9] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," *IEEE Automatic Speech Recognition and Understanding Workshop*, December 2003.

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

[11] J. Zheng and A. Stolke, "Improved discriminative training using phone lattices," *InterSpeech*, September 2005.

[12] H. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," *ICASSP*, 2000.