

# THE BBN BYBLOS 2000 CONVERSATIONAL MANDARIN LVCSR SYSTEM

*Han Shu, Chuck Wooters, Owen Kimball,  
Thomas Colthurst, Fred Richardson, Spyros Matsoukas, Herbert Gish*

GTE/BBN Technologies  
Cambridge, MA 02138  
hshu@bbn.com

## ABSTRACT

This paper describes the year 2000 BBN Byblos Mandarin large vocabulary conversational speech recognition (LVCSR) system, the winning (and only) Mandarin system from the Spring 2000 Hub-5 evaluation sponsored by NIST. We first outline the training and decoding procedures used in the system, and describe the performance of the system used in the evaluation. We then describe the effect of several features that were not in the evaluation system but have been added since, including Jacobian compensated Vocal Tract Length Normalization (VTLN), system combination, a higher number of system parameters, and additional training data. Together these give an additional 5.4% relative improvement on character error rate (CER) from the evaluation system.

## 1. INTRODUCTION

This paper describes the BBN Byblos Mandarin system that was entered in the Spring 2000 NIST LVCSR evaluation. The evaluation test consisted of twenty 5-minute conversations of fluent Mandarin taken from the CallHome database. The BBN system was the only system entered in the Mandarin evaluation and achieved a character error rate (CER) of 57.1% on the Eval2000 test set.

BBN's Mandarin system was developed for this evaluation in the relatively short time of about 7 weeks. In the last Hub-5 Mandarin evaluation in 1997, the Department of Defense submitted the winning system [1], which was a version of BBN's Byblos recognition system that included a number of refinements introduced by DOD researchers. These refinements included an optimized phoneme set, a better phonetic dictionary with tone specifications, and more language modeling data. Although built using Byblos, this system was unfortunately unavailable to us in the time we had to develop this year's system. This year's system was developed from a significantly older baseline system, run by BBN five years ago for the 1995 Hub-5 Mandarin evaluation, which lacked a number of features we have introduced into Byblos recognition system since then.

Fortunately, the development of this year's system was simplified by the general philosophy of using language-independent technologies wherever possible. Specifically, most of the development was focused on integrating and testing the features that have proven most successful in BBN's English version of the Byblos system. We have found that this approach not only saves

development time but that most improvements to our speech recognizer are useful across languages.

There were a number of features that were not incorporated in the evaluation system because of a lack of time but that were added to the system after the evaluation. These changes yielded an additional 5.4% relative improvement from the submitted system's CER.

The outline of this paper is as follows. We first give an overview of the task, including a review of the properties of the Mandarin language and a description of the evaluation data set. We then describe the Byblos recognition system and the specific configuration of the system at the time of the evaluation. Finally, we describe experiments involving a number of improvements made to the system after the evaluation.

## 2. TASK DESCRIPTION

### 2.1. Fundamentals of Mandarin

Mandarin is the standard dialect of Chinese. Unlike English, Chinese is character-based and words are not well-defined units. Chinese words consist of either one character alone or compound words consisting of two or more characters. The word boundaries are ambiguous, and the word boundaries are not customarily marked in written Chinese text.

Each character's pronunciation consists of a single syllable and each syllable in turn consists of an initial consonant, a medial vowel, a central vowel, and a syllabic ending, where the initial consonant, the medial vowel, and the syllabic ending are optional. There are 24 initial consonants, 4 medial vowels, 13 central vowels, and 4 syllabic endings in total. Tone, the movement of pitch, also plays a major role in Chinese. Each character can have one of five tones and there are sets of characters that can be distinguished from each other only by their tone. With more than 6,000 frequently used characters, and only approximately 1,300 tone-specific syllables (approximately 400 non-tone-specific syllables), each character can have many homophones.

### 2.2. Challenges of the Task

In addition to the usual challenges of recognizing fluent conversational speech, and the unique characteristics of Mandarin

described in the last section, several other factors contribute to the difficulty of this task. First, the amount of the acoustic training data is small, consisting of 100 CallHome conversations, or about 15 hours of speech. Second, the amount of language model training data is also small, consisting of the same 100 CallHome conversations used for acoustic model training plus 42 CallFriend transcriptions. An n-gram grammar on the character level would have a much higher perplexity than a n-gram grammar on the word level. The higher perplexity along with the high number of homophones on the character level would result in a much higher CER. Thus, we have chosen to use a grammar that is on the word level. The 100 CallHome conversation transcriptions are segmented on the word level. Because of the ambiguous nature of word boundaries in Chinese, it was difficult to obtain additional transcriptions similarly segmented on the word level for training the language model. Third, the CallHome conversations are international telephone calls recorded over the telephone line. Crosstalk and background noises such as babies crying and car noises can be heard in the recording. Lastly, the telephone callers vary significantly in their background, the pronunciation variation and the different accents of the callers all increase the difficulty of this Mandarin task.

## 2.3 Data Set

All the systems described in this paper were trained on the CallHome training set or a combination of the CallHome and CallFriend training sets. The CallHome training set consists of 100 CallHome conversations, or a total of 15 hours of speech; the CallFriend training set consists of 42 CallFriend conversations, or a total of 20 hours of speech (CallFriend conversation segments are typically longer than CallHome).

The systems were tested on the 1995 (Eval95), 1997 (Eval97), and 2000 (Eval2000) CallHome Mandarin evaluation sets. These test sets each contain 20 different telephone conversations, with each conversation containing about five minutes of speech.

# 3. EVALUATION SYSTEM DESCRIPTION

## 3.1. Signal Processing

The 2000 BBN Mandarin LVCSR system uses a single, 45-dimensional feature stream. Features are extracted from overlapping frames of audio data, each 25ms long, at a rate of 100 frames per second. Each frame is windowed with a Hamming window, and then an LPC-smoothed, Vocal Tract Length Normalization (VTLN) warped log power spectrum is computed for the frequency band 125-3750 Hz. From this spectrum, 14 Mel-warped cepstral coefficients are computed. We use a gender-independent, 128 term Gaussian mixture model to compute a maximum-likelihood VTLN warp parameter [2,3]. (In the evaluation system, the VTLN warp was estimated using an older method that did not compensate for the Jacobian of the VTLN transformation; the effect of adding this compensation is investigated in Section 4.) The Mandarin evaluation system was gender-independent so no gender detection calculation is performed. The mean cepstrum and peak energy of each

conversation is removed non-causally from the appropriate sub-vectors. In addition, the feature vectors are scaled and translated so that, for each conversation side, each cepstral feature has zero mean and unit variance. These 14 base cepstral features and the frame energy, together with their first and second derivatives, compose the final 45-dimensional feature vector. We have not yet incorporated pitch in our signal processing, although we expect this to help performance given the tonal nature of Mandarin.

## 3.2. Acoustic and Language Model Training

The acoustic training for the BBN Mandarin system builds two sets of gender-independent models, phoneme tied mixture (PTM) and state clustered tied mixture (SCTM) models, which are used in the different passes of the Byblos recognizer. The training set used for the evaluation was the 15-hour CallHome training set (section 4 describes the effect of training with a larger data set). The training data is first labeled using a forced phonetic alignment to simple bootstrapped models (64-Gaussian, PTM models trained from flat initial estimates). The labels are then used to grow separate decision tree for both the Gaussian clusters and their mixture weights clusters.\* The five state Hidden Markov Model (HMM) transition probabilities are unclustered.

Following clustering, the Gaussians for the final models are initialized via the k-means algorithm, and finally, all the parameters of the models are trained with three passes of the EM algorithm. This process is done for both the crossword SCTM and the non-crossword PTM. For English recognition, where larger training sets are available, we typically model contexts using quinphones (i.e. the preceding and following two phonetic contexts of a phone), but with this relatively small Mandarin training set, both SCTM and PTM models use triphone context only. The coarse PTM models use approximately 22,800 Gaussians (89 phonemes with 256 Gaussians each) and 6,000 mixture weight clusters, while the fine SCTM models use 32,000 Gaussians (1000 state clusters with 32 Gaussians each) and 12,000 mixture weight clusters.

The phoneme set consists of 89 tone-specific phones, and the dictionary contains 11,600 words composed from 2191 individual characters. The trigram language model is trained on transcriptions from 100 CallHome and 42 CallFriend conversations, approximately 632,000 words in total.

## 3.3. Recognition

---

\* For each state in a triphone HMM, there are two levels of parameter sharing. The first level specifies the sharing of the Gaussians among triphones of the same state, the second level specifies the sharing of mixture weights among triphones of the same state. The Gaussian clustering tree is a subtree of the mixture weight tree so that different distributions can share the same set of Gaussians using separate weights. We call the first level sharing "Gaussian clusters", and the second level sharing "mixture weight clusters".

Decoding is performed in two stages: the first stage uses speaker independent models, the second stage uses MLLR speaker-adapted models adapted to the recognition result from the first stage. The evaluation system did not use system combination.

Both the unadapted and adapted stages of decoding each use a multi-pass recognizer [4, 5] that operates as follows: the first pass is a forward fast match that uses non-crossword PTM models and a bigram language model. The second and third passes perform backward and forward searches respectively, both using the same PTM acoustic models but with an approximate trigram language model. Times and scores for word starts and ends are saved in these passes and from this information a word lattice is created. The next pass of the recognizer searches this lattice with crossword SCTM acoustic models and a trigram language model; it produces an N-best list of the top 100 ranked possible transcriptions. The N-best list is finally reordered using optimized weights to get a single best hypothesis. The evaluation system also computed confidence scores using features generated on the adapted stage's N-best output.

### 3.4. Evaluation System Performance

Table 1 summarizes the unadapted and adapted recognition performance of the BBN Mandarin system in terms of character error rate (CER) on a development set (Eval97, originally the test set for the 1997 CallHome evaluation) and on the 2000 evaluation data (Eval2000); the evaluation system achieved 54.% CER on the Eval97 test set and 57.1% on the Eval2000 test set. For comparison, the winning system for the last Hub-5 Mandarin evaluation in 1997 achieved 53.8% on the Eval97 test set. That system differed from the BBN 2000 system in that it included pitch, language model training data from broadcast domain, a simpler VTLN system, and a larger lexicon with 27,600 words. Given our limited development time, we were pleased to have achieved essentially the same performance on that test set. In the next section we describe changes to the system made shortly after the evaluation that significantly improved our Mandarin system from this point.

Test Set	Unadapted	Adapted
Eval97	57.1%	54.0%
Eval2000	60.3%	57.1%

**Table 1:** Performance of the BBN 2000 Evaluation System on the Eval97 test set and the Eval2000 test set.

## 4. EXPERIMENTS IN LVCSR

### 4.1. Increased Number of Parameters

In the evaluation system, the coarse PTM models use approximately 22,800 Gaussians (89 phonemes with 256 Gaussians each) and 6,000 mixture weight clusters, while the fine SCTM models use 32,000 Gaussians (1000 state clusters with 32 Gaussians each) and 12,000 mixture weight clusters. We suspected that we did not have enough parameters in our system.

The system contains thresholds that control the number of Gaussian clusters and the number of mixture weight clusters based on the amount of training data for each triphone. By relaxing these thresholds and increasing the number of Gaussians per cluster, we can add more parameters to the system. In the new system, the coarse PTM models use approximately 22,800 Gaussians (89 phonemes with 256 Gaussians each) and 7,800 mixture weight clusters, while the fine SCTM models use 76,800 Gaussians (1,200 state clusters with 64 Gaussians each) and 21,700 mixture weight clusters. By increasing the number of parameters in the system, we obtained a 0.8% absolute reduction in CER. Table 2 summarizes the effect of increasing the number of parameters.

Number of Parameters		CER
PTM	SCTM	
22,800 Gaussians 6,000 mixture weight clusters	32,000 Gaussians 12,000 mixture weight clusters	62.6%
22,800 Gaussians 7,800 mixture weight clusters	76,800 Gaussians 21,700 mixture weight clusters	61.8%

**Table 2:** The effect of increasing the number of parameters on the Eval95 test set.

### 4.2. Additional Training Data

The training data used in the evaluation system consists of 100 CallHome conversations, roughly 15 hours of speech. By adding 42 CallFriend conversations, we increased the amount of training speech to 35 hours. Table 3 summarizes the effect of adding the CallFriend conversations: we observe a 1.2% absolute reduction in error on the Eval95 test set.

Training Data for System	Total training (hours)	CER
100 CallHome conversations	15	61.8%
+ 42 CallFriend conversations	35	60.6%

**Table 3:** The effect of increasing acoustic training data on the Eval95 test set.

### 4.3. Improved VTLN with Jacobian Compensation

VTLN attempts to normalize the cepstral feature variability due to different vocal tract lengths among speakers. In the evaluation system we used a maximum-likelihood VTLN (ML-VTLN) procedure that was developed several years ago [2,3]. The ML-VTLN approach uses a Gaussian mixture model (GMM) against which speakers are scored at a multiplicity of warps. The warp that scores the highest likelihood is then taken to be the VTLN stretch factor for that speaker. One deficiency of this approach is that the GMM shows an inherent likelihood bias for cepstra at different warps. To compensate for this effect, the determinant of

the VTLN transformation is estimated empirically per speaker and applied to remove this bias. Table 4 shows the results for adding this feature: the empirical Jacobian compensation yields a 0.9% absolute reduction in error.

System	CER
VTLN <i>without</i> Jacobian compensation	60.6%
VTLN <i>with</i> Jacobian compensation	59.7%

**Table 4:** CERs for VTLN with and without Jacobian compensation on the Eval95 test set.

To validate the improvements shown in sections 3.1, 3.2, and 3.3, we tested a system with all three improvements on a second test set, the Eval97 test set. Specifically, we ran a system that included more parameters, more training data, and the improved VTLN. Table 5 shows the improvement due to these changes for both the unadapted and adapted recognition results.

System on Eval97	Unadapted	Adapted
Evaluation System	57.1%	54.0%
+ more parameters + additional training data + Improved VTLN	54.6%	51.6%

**Table 5:** Comparison of performance between the baseline evaluation system and improved system on the Eval97 test set.

#### 4.4. System Combination

Due to lack of time, the evaluation system did not include system combination. To run system combination, the lattice scoring and N-best reordering passes of the recognizer’s adapted stage are run two additional times after adaptation, using the same adapted models but taking as input cepstra calculated at different frame rates, 80 and 125 frames per second. Character-level confidences were calculated for each of these systems using a generalized linear model (GLM). The major features for confidence selected by the GLM in training include: frequency of occurrence in the 100-best list, word duration, normalized utterance acoustic score, and the number of hypothesized silences for the utterance. Following individual confidence generation, the results from all three frame rates were combined using a modified ROVER method [6], in which the vote for a hypothesized character is weighted using the systems’ input character confidences. Confidences on the character level for this final combined system are again calculated using a GLM. The major features used for calculating the final combined system confidence include: ROVER score, individual system confidences, and word duration.

Table 6 summarizes the character error rate and the normalized cross entropy (NCE) for the three systems at different frame rates and their combined system. Using system combination in this way gives a 0.5% absolute reduction in CER from the baseline 100 frames per second system. The final combined system achieved a better NCE score for its character-level confidences as well.

System	CER	NCE
80 frames/second system	53.9%	0.168
100 frames/second system	51.6%	0.168
125 frames/second system	51.9%	0.156
<b>Final combined system via ROVER-voting</b>	<b>51.1%</b>	<b>0.179</b>

**Table 6:** Performance of system combination with modified ROVER on the Eval97 test set.

#### 4.5. Summary

This paper has described the BBN Mandarin system that was used in the NIST 2000 Mandarin evaluation. We described a system that was rapidly developed and that relied primarily on language-independent features. Improvements from the evaluation system were obtained by increasing the number of parameters in the system, adding training data, improving VTLN with Jacobian compensation, and using system combination. Together these improvements achieved a 51.1% CER on the Eval97 test data, a 2.9% absolute or 5.4% relative improvement from the baseline evaluation system of 54.0% CER. There are a number of other possible improvements that we plan to incorporate in the future, including adding more language modeling data, adding pitch information, and investigating the use of character n-gram language models in combination with word n-grams to help improve the character error rate on out of vocabulary words.

### 5. REFERENCES

1. Caetano, P.V., “Porting LVCSR to Mandarin”, *Presentation at LVCSR Workshop*, November, 1997.
2. S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, “Speaker normalization on conversational telephone speech”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 1996.
3. G. Zavaliagos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, and H. Gish, “The BBN Byblos 1997 large vocabulary conversational speech recognition system”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, May 1998.
4. L. Nguyen and R. Schwartz, “Single-Tree Method for Grammar-Directed Search”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 613-616, Phoenix, AZ., March 1997.
5. L. Nguyen and R. Schwartz, “Efficient 2-Pass N-Best Decoder”, *Proceedings of EuroSpeech*, pp. 167-170, Rhodes, Greece, Sept. 1997.
6. J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)”, *Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347-354, Santa Barbara, 1997.