RECENT ADVANCES IN THE IBM GALE MANDARIN TRANSCRIPTION SYSTEM[1]

Stephen M. Chu, Hong-kwang Kuo, Lidia Mangu *Yi Liu, Yong Qin, Qin Shi, Shi Lei Zhang* *Hagai Aronowitz*[2]

IBM T. J. Watson Research Center
{schu, hkuo, mangu}@us.ibm.com

IBM China Research Lab
{liuyyi, qinyong, shiqin, slzhang}@cn.ibm.com

IBM Haifa Research Lab
hagaia@il.ibm.com

ABSTRACT

This paper describes the system and algorithmic developments in the automatic transcription of Mandarin broadcast speech made at IBM in the second year of the DARPA GALE program. Technical advances over our previous system include improved acoustic models using embedded tone modeling, and a new topic-adaptive language model (LM) rescoring technique based on dynamically generated LMs. We present results on three community-defined test sets designed to cover both the broadcast news and the broadcast conversation domain. It is shown that our new baseline system attains a 15.4% relative reduction in character error rate compared with our previous GALE evaluation system. And a further 13.6% improvement over the baseline is achieved with the two described techniques.

Index Terms **–** speech recognition, speech processing, tone modeling, topic adaptation.

1. INTRODUCTION

This paper describes Mandarin speech recognition technology developed at IBM for the Global Autonomous Language Exploitation (GALE) program. The overall goal of this program is to extract information from publicly available broadcast sources in multiple languages, and to make it accessible to monolingual English speakers. In order to accomplish this, the program has several major components: *speech recognition*, *machine translation*, and *question answering*. The focus of this paper is on the Mandarin language automatic speech recognition (ASR) component.

The GALE program focuses on two types of broadcast audio: broadcast news – which was a focus of attention in the previous DARPA Effective Affordable Reusable Speech-to-text (EARS) and HUB-4 programs – and broadcast conversations. The study of broadcast conversations is relatively new to the speech recognition community, and the material is more challenging than broadcast news shows. Whereas broadcast news material usually includes a large amount of carefully enunciated speech from anchor speakers and trained reporters, broadcast conversations are less scripted and more spontaneous in nature, with the associated problems of spontaneous speech: pronunciation variability, rate-of-speech variability, mistakes, corrections, and other disfluencies.

This paper focuses on our progress in Mandarin recognition made during the second year of the GALE program. In addition to engineering refinements in system building, two technical improvements that have led to clear performance gains are discussed: (*a*) embedded tone modeling in the acoustic models, and (*b*) dynamic topic adaptation for language model (LM) rescoring.

The remainder of this paper is organized as follows. In Section 2, we give an overview of our system architecture, and discuss the specifics of our acoustic modeling pipeline. Section 3 and 4 describe the tone modeling and the topic-adaptive language model rescoring techniques, respectively. Section 5 covers the experimental setup and results, followed by conclusions in Section 6.

2. SYSTEM ARCHITECTURE

The IBM GALE Mandarin broadcast speech transcription system operates in multiple passes. Fig. 1 shows the processing pipeline from the perspective of acoustic modeling.

2.1. Front-End Processing

The basic features used for segmentation and recognition are *perceptual linear prediction* (PLP) features. Feature mean normalization is applied as follows: in segmentation and speaker clustering, the mean of the entire session is computed and subtracted; for SI decoding, speaker-level mean normalization is performed based on the speaker clustering output; and at SA stage, the features are mean and variance normalized for each speaker. Consecutive feature frames are spliced and then projected back to a lower dimensional space using *linear discriminant analysis* (LDA), which is followed by a *maximum likelihood linear transform* (MLLT) [2] step to further condition the feature space for diagonal covariance Gaussian densities.

2.2. Speaker Diarization

The speech segmentation is done using Viterbi decoding of a state-loop containing single-state speech and non-speech HMMs. The speech model is created by clustering the Gaussians of the speech triphones. The non-speech model is similarly created by clustering the Gaussians of the non-speech triphones. Interestingly, no degradation in performance was observed when using Arabic acoustic models compared to using Mandarin acoustic models for segmentation.

The resulting speech-labeled segments are clustered using the *K*-means algorithm. The number of clusters is first set to achieve a mean cluster size of 30sec (but no more than 20 clusters). Then, the speech input is divided linearly to form the requested number of clusters. Each cluster is modeled by a Gaussian with a diagonal

[1] This work was funded in part by DARPA contract HR0011-06-2-0001.

[2] The author was at IBM T. J. Watson Research Center during this work.

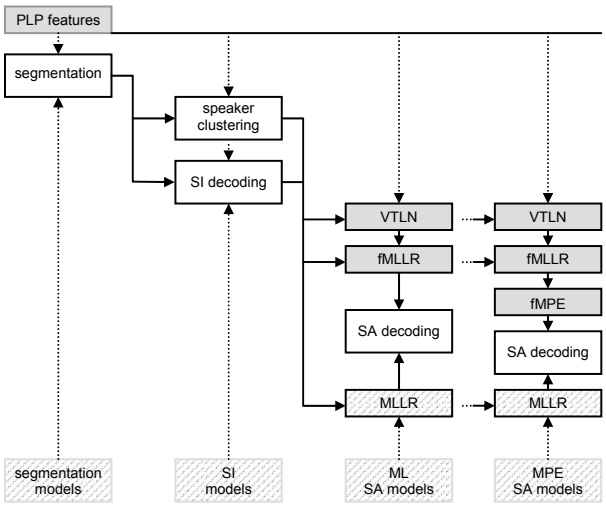

Fig. 1. The IBM Mandarin broadcast speech transcription system consists of speech segmentation, speaker clustering, speaker independent (SI) decoding, and speaker adapted (SA) decoding.

Covariance. After initialization, 5 iterations of *K*-means are executed using the *Mahalonobis* distance. Finally, the resulting clusters are merged by selection of the closest pair of clusters until a stopping criterion of a maximal distance of 0.2 is reached.

For post processing, short silences (< 0.3sec) are eliminated if the neighboring speech segments are labeled as the same cluster.

2.3. SI Models

Phones are represented as three-state, left-to-right HMMs. With the exception of silence and noise, the HMM states are context-dependent conditioned on quinphone context covering both past and future words. The context-dependent states are clustered into equivalence classes using a decision tree.

Emission distributions of the states are modeled using mixtures of diagonal-covariance Gaussian densities. The allocation of mixture component to a given state is a function of the number of frames aligned to that state in the training data. Maximum likelihood (ML) training is initialized with state-level alignment of the training data given by an existing system. A mixture-splitting algorithm iteratively grows the acoustic model from one component per state to its full size. One iteration of Viterbi training on word graphs is applied at the end.

2.4. SA Models

The SA acoustic models share the same basic topology with the SI model. For speaker adaptation, a model-space method, *maximum likelihood linear regression* (MLLR), and two feature-space methods, *vocal tract length normalization* (VTLN) and *feature-space* MLLR (fMLLR) [3], are used in the baseline system.

An eight-level binary regression tree is used for MLLR, which is grown by successively splitting the nodes from the top using soft K-means algorithm. The VTLN frequency warping consists of a pool of 21 piecewise linear functions, or warping factors. In decoding, a warping factor is chosen such that it maximizes the likelihood of the observations given a voice model built on static features with full covariance Gaussian densities.

In addition to the speaker adaptation procedures, the improved Mandarin transcription system also employs the discriminately trained *minimum phone error* (MPE) [4] models and the *feature-space* MPE (fMPE) [5] transform.

3. TONE MODELING

Treatments of tones in Mandarin ASR can be grouped into two categories: explicit tone recognition and embedded tone modeling. The latter approach, considered in this work, inherently allows joint optimization of tonal/atonal parameters in a coherent system by incorporating tonal components in both the acoustic models and the feature stream.

3. 1. Tonal Phonetic Units

A straightforward tonal phone set can be derived directly from the *Pinyin* system, which maps Mandarin phonetics using a set of *initials* and *finals*. It we treat each initial and each tonal final as phonemes, the total number of phonemes is $N = C + T \times V$, where C is the number of consonant initials, T is the number of tones, and V is the number of vowel endings. This would lead to a large set of $20 + 5 \times 39 = 215$ phones.

By grouping the glides with the consonant initials into *premes*, each syllable can be decomposed into *demisyllables*. With the demisyllable system, N is reduced to: $53 + 5 \times 21 = 158$, where 53 is the number of premes and 21 is the number of the second demisyllable without glide. Adding four auxiliary phonemes (undefined, word boundary, intra-word silence, and inter-utterance silence) brings the total number of phonemes in the set to 162. This tonal demisyllable system is used as our basic phonetic units for all systems described in this paper.

Note that the number of phonemes can be drastically reduced if we assume that the pitch information on the main vowel is sufficient to determine the tone of a whole syllable. We compared a system built on the resulting 77-phone set and the demisyllable system, and found the latter to be superior with large training sets, thus more suitable to the GALE scenario.

3.2. Tonal Features

The frontend described in Section 2.1, which is used in our baseline system, does not capture pitch information in speech. The tonal system augments the standard frontend with explicit pitch features extracted using the YIN [6] algorithm.

The pitch tracker gives two outputs: an estimated pitch contour of the audio signal, and a frame-level confidence score given by the algorithm, as shown in Fig. 2(*a*). The pitch is undefined in unvoiced region and during silences, and can thus cause large noise spikes in the estimation, which, if left untreated, will have an adverse effect on statistical modeling. To address this problem, we first use the confidence score to classify voiced and unvoiced regions of the speech signal. Then the pitch values in the unvoiced segments are replaced through *piecewise cubic Hermite interpolating polynomial* using samples in the adjacent voiced regions. The

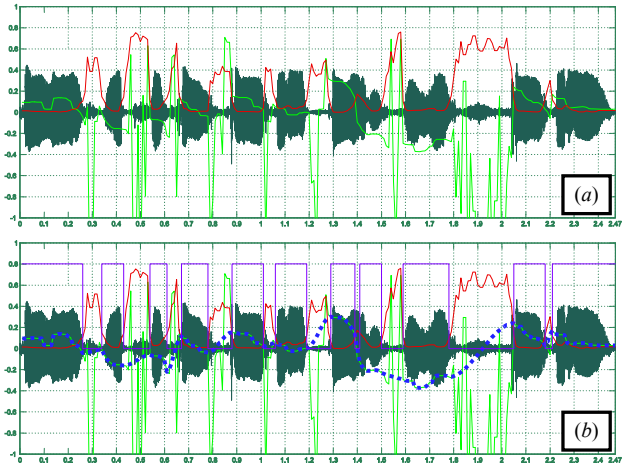

Fig. 2. (*a*) Pitch tracker output, where the green line is the estimated pitch value and the red line is the confidence score (lower is more confident). (b) Interpolation and smoothing results. The square signal shows the voiced/unvoiced classification, and the dotted line is the pitch feature after interpolation.

procedure and the result are demonstrated in Fig. 2(*b*). Further smoothing and normalization are applied using median and moving average filtering.

Because the confidence score provides strong cues for voicing, we also add it as an additional feature dimension; together with the smoothed pitch contour, a two-dimensional tonal feature vector is thus constructed.

3.3. Feature Fusion

The five tones in Mandarin are defined by relative pitch movement rather than absolute frequency value. Therefore, the raw pitch features should be processed such that the underlying dynamics can be captured.

One way is to fuse the tonal features with the standard PLP features *early*, so that the subsequent splicing and LDA transforms can extract the dynamics in a data-driven manner. The other is to compute the Δ and $\Delta\Delta$ of the pitch first and integrate them with the regular feature stream *late* after the LDA projection. We implemented both methods and found that the *early* fusion approach gained more from the tonal features. Results on the tonal system reported in this paper are from the first implementation.

4. TOPIC-ADAPTIVE LM RESCORING

During the first pass decoding, the speech recognizer generates a word hypothesis lattice. This lattice can be further processed by LM rescoring to improve the speech recognition accuracy. For example, a larger LM is often used or an LM adapted to the first pass decoded output. To improve the effectiveness of LM rescoring, we developed two methods which will be described in further detail below. The first method is called topic-based rescoring, and the second is called dynamic topic LM. The LMs from both methods are interpolated with the general LM in a snippet-specific manner during LM rescoring.

4.1. Static Topic LM

The idea behind the first method, topic-based language modeling, is to create a number of homogeneous LMs that cover specific topics. In particular, more than 20,000 Chinese news articles were collected and annotated to build an SVM classifier based on the *structural risk minimization* (SRM) principle. The raw feature to represent each training sample is a vector of terms given by the IBM CRL Chinese text segmenter. Using this classifier, our LM training corpus is organized into a tree structure with 129 leaf nodes. An on-topic LM is trained for each of the 129 classes. We refer to these as *static topic LMs*.

4.2. Dynamic Topic LM

In contrast to fixed topics in the first method, the second method, dynamic topic LM, flexibly assembles text from the pool of data in the GALE collection to model the topic of each snippet.

To efficiently implement this idea, we leverage an open-source text search engine (*Lucene*) to dynamically collect the most relevant documents over our 5G LM training corpus. The ranking algorithm used here is a combination of *vector space model* (VSM) of information retrieval and the Boolean model to determine how relevant a given document is.

First, the 1-best hypothesis from the first pass decoding for each snippet is chopped into short phrases according to the significant silence words and hesitation words. Then, after deleting every short phrase whose length is shorter than a predefined threshold, we obtain a certain number of search keys for that snippet. For each search key, we sort the retrieved documents according to the ranking score; top n documents with the highest scores are kept. Finally, for all the search keys, we sum up the final score for each document which indicates how many times the document is hit by those search keys and this count information is the final criteria determining its relevance.

In the experiments, we keep the top 20 documents according to the hit score and build a small 4-gram model for each snippet. We call this the *dynamic topic LM*.

As shown in Fig. 3, in the decoding phase, the general LM is first used to generate a word-lattice and 1-best hypothesis. For each snippet, the 1-best hypothesis is used to choose 3 static topic

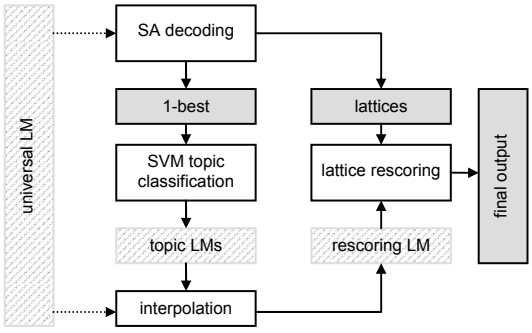

Fig. 3. Topic adaptation is carried out through lattice rescoring with an LM interpolated from the universal LM and 4 topic-specific LMs. Topic classification is based on the 1-best word hypothesis given by the SA decoding output.

LMs and also used to build a dynamic topic LM. We then interpolate the general LM, 3 static topic LMs and 1 dynamic topic LM together to minimize the perplexity of the first pass decoded output, and obtain a final LM for LM rescoring.

5. EXPERIMENTS

5.1. System Setup

For the baseline system, the 16 KHz input signal is coded using 13-dimensional PLP features with a 25ms window and 10ms frame-shift; 9 consecutive frames are spliced and projected to 40 dimensions using LDA. For the tonal system, the 2-dimensional tonal features are appended to the PLP features to form a 15-dimensional feature vector, and the rest of the processing pipeline is unchanged.

All acoustic model training are based on 1,321 hours (1,055 hours fully transcribed, 266 hours lightly-supervised [7]) of acoustic training data released by LDC for the GALE program. The SI acoustic model has 10K quinphone states modeled by 300K Gaussian densities. The SA model uses a larger tree with 15K states and 500K Gaussians.

The decoding LM is built by interpolating 20 back-off 4-gram models using modified Kneser-Ney smoothing. The interpolation weights are chosen to optimize the perplexity of a 364K held-out set. In total, 5GB of text data is used in training. The final LM has 6.1M n-grams and 107K words.

5.2 Experimental Results

Three test sets are used in the experiments. These sets contain Mandarin TV broadcast collected from stations in mainland China, Taiwan, and Hong Kong. The first is the evaluation set from the 2006 GALE evaluation, referred to as *eval*'06 here. It contains 63 minutes of audio, and has 18.3K characters in the reference transcript. The second test set, denoted *dev*'07, is defined by the GALE participants. This set contains 2 hours and 32 minutes of audio and 44.6K characters in the reference. The third set *eval*'07 is the evaluation set from the 2007 GALE evaluation, which contains 2 hours and 21 minutes of audio and 40.6K characters in the reference. Results on *eval*'07 are further divided into broadcast news (*bn*) and broadcast conversations (*bc*).

The performance of the baseline system and the tonal system is compared in Table 1. The tonal system shows clear improvement over the baseline from SI level decoding to MPE/fMPE SA models with fMLLR and MLLR adaptations. Further reduction in CER is achieved when the tonal MPE/fMPE system is cross-adapted on the baseline decoding output, as shown in Table 2.

Table 1. Comparing CER of the baseline system (BASE) and the tonal system (TONE) on *eval*'06 and *dev*'07.

			BASE	TONE
SI:	----	*eval*'06:	24.6	23.5
	----	*dev*'07:	18.9	18.5
SA:ML	+fMLLR	*eval*'06:	21.9	21.3
		dev'07:	15.7	15.4
	+MLLR	*eval*'06:	21.7	21.2
		dev'07:	15.4	15.0
SA:MPE	+fMLLR	*eval*'06:	18.6	18.1
		dev'07:	12.4	12.0
	+MLLR	*eval*'06:	18.5	17.9
		dev'07:	12.2	11.8

Table 2. Tonal system cross-adapted on the baseline system (BASE × TONE) reduces CER further. Topic-adaptive LM rescoring brings substantial gains.

	dev'07	*eval*'06	*eval*'07	*bn*, *bc*
BASE:	12.2	18.5	11.0	4.9, 18.6
TONE:	11.8	17.9	10.6	4.5, 18.1
BASE × TONE:	11.5	17.6	10.2	4.4, 17.4
Topic LM + BASE × TONE:	10.9	17.1	9.5	3.6, 16.8

Compared with our 2006 GALE evaluation system, which has CER of 13.0% on *eval*'07, the new baseline's 11.0% CER represents a 15.4% relative reduction in errors. Tone modeling and the topic-adaptive LM rescoring technique further reduce the CER by 0.8 and 0.7 absolute points, respectively, from the new baseline, together they achieve 13.6% relative error reduction.

6. CONCLUSIONS

This paper considers the Mandarin broadcast speech transcription task in the context of the DARPA GALE project. We present a state-of-the-art Mandarin ASR architecture and described the technical improvements made in the 2007 evaluation system.

REFERENCES

[1] S. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcriptions at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1569-1608, September 2006.

[2] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximimum likelihood discriminant feature spaces," in *Proc. ICASSP'00*, vol. 2, pp. 1129-1132, June 2000.

[3] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75-98, April 1998.

[4] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP'02*, May 2002.

[5] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: discriminatively trained features for speech recognition," in *Proc. ICASSP'05*, vol. 1, pp. 961-964, March 2005.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, pp. 1917-1930, 2002.

[7] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP'04*, vol. 1, pp. 737-740, May 2004.

[8] M. J. F. Gales, A. Liu, K. C. Sim, P. C. Woodland, and K. Yu, "A Mandarin STT system with dual Mandarin-English output," presented at GALE PI Meeting, Boston, March 2006.

[9] B. Xiang, L.Nguyen, X. Guo, and D. Xu, "The BBN mandarin broadcast news transcription system," in *Proc. Interspeech'05*, pp. 1649-1652, September 2005.

[10] R. Sinha, M. J. F. Gales, D. Y. Kim, X. A. Liu, K. C. Sim, P. C. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *Proc. ICASSP'06*, May 2006.