THE IBM MANDARIN BROADCAST SPEECH TRANSCRIPTION SYSTEM

Stephen M. Chu[1]*, Hong-kwang Kuo*[1]*, Yi Y, Liu*[2]*, Yong Qin*[2]*, Qin Shi*[2]*, and Geoffrey Zweig*[1]

[1]IBM T. J. Watson Research Center
{schu, hkuo, gzweig}@us.ibm.com

[2]IBM China Research Lab
{liuyyi, qinyong, shiqin}@cn.ibm.com

ABSTRACT

This paper describes the technical and system building advances in the automatic transcription of Mandarin broadcast speech made at IBM in the first year of the DARPA GALE program. In particular, we discuss the application of *minimum phone error* (MPE) discriminative training and a new topic-adaptive language modeling technique. We present results on both the RT04 evaluation data and two larger community-defined test sets designed to cover both the broadcast news and the broadcast conversation domain. It is shown that with the described advances, the new transcription system achieves a 26.3% relative reduction in character error rate over our previous best-performing system, and is competitive with published numbers on these datasets.

Index Terms **–** speech recognition, speech processing.

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 more unplanned and spontaneous in nature, with the associated problems of spontaneous speech: pronunciation variability, rate-of-speech variability, mistakes, corrections, and other disfluencies.

This paper will describe our Mandarin recognition work from the system-building perspective. The main contributions of the paper are: (*a*) the presentation and validation of an effective system architecture, (*b*) the application of the MPE/fMPE algorithm in our Mandarin system, and (*c*) an adaptive language modeling technique. The remainder of this paper is organized as follows. In Section 2, we present our system architecture. This architecture amalgamates techniques used previously in English [1], as well as extending it with a novel adaptive language modeling technique. In Section 3, we describe the specifics of our Mandarin system, including the training data and system size. Section 4 presents experimental results on broadcast news and broadcast conversation test sets, followed by conclusions in Section 5.

2. SYSTEM ARCHITECTURE

The IBM GALE Mandarin broadcast speech transcription system is composed of three main stages, speech segmentation/speaker clustering, speaker independent (SI) decoding, and speaker adapted (SA) decoding. A system diagram is shown in Fig. 1.

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. Segmentation and Clustering

The segmentation step uses an HMM-based classifier. The speech and non-speech segments are each modeled by a five-state, left-to-right HMM with no skip states. The output distributions are tied across all states within the HMM, and are specified by a mixture of Gaussian densities with diagonal covariance matrices.

After segmentation, the frames classified as non-speech are discarded, and the remaining segments are put through the clustering procedure to give speaker hypotheses. The clustering algorithm models each segment with a single Gaussian density and clusters them into a pre-specified number of clusters using K-means.

Note that in the broadcast scenario, it is common to observe recurring speakers in different recording sessions, e.g., the anchors of a news program. Therefore, it is possible to create speaker clusters beyond the immediate broadcast session. Nevertheless, in the scope of this paper, we shall restrict the speaker clustering procedure to a per session basis.

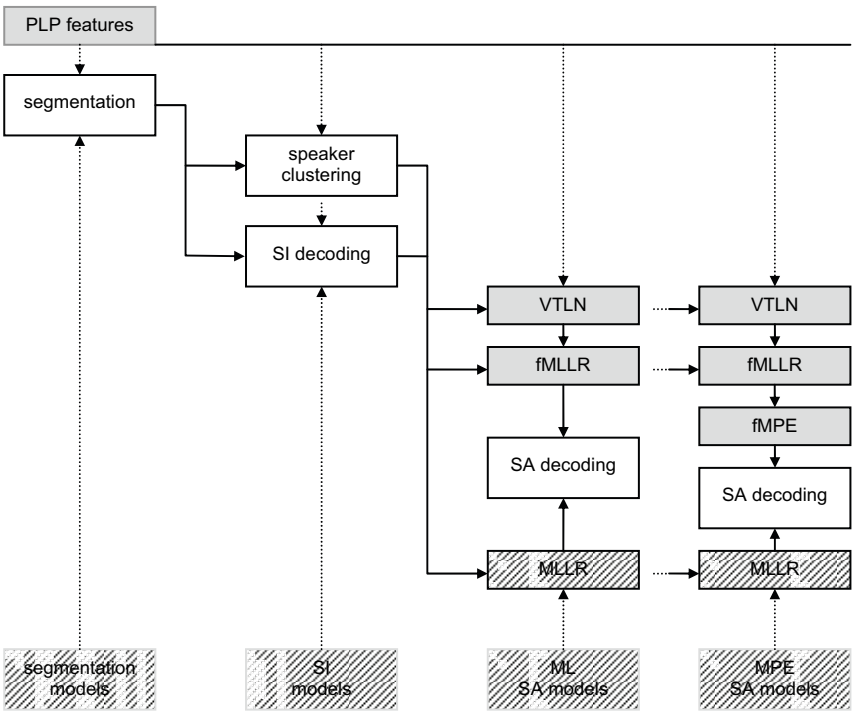

Fig. 1. The IBM Mandarin broadcast speech transcription system consists of speech detection/segmentation, speaker clustering, speaker independent decoding, and speaker adapted decoding. In speaker adapted decoding, both feature and model space adaptations are applied. Models and transforms discriminatively that are trained using *minimum phone error* training provide further refinement in acoustic modeling.

2.3. SI Models

The system uses a tone-specific phone set with 162 phonemes. 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) [3] and *feature-space* MLLR (fMLLR) [4], 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) [5] models and the recently developed *feature-space* MPE (fMPE) [6] transform. Experiments show that these discriminative algorithms give a significant improvement to recognition performance.

2.5. Language Modeling

The language models (LM) considered in this work are interpolated back-off 4-gram models smoothed using modified Kneser-Ney smoothing [7]. The interpolation weights are chosen to optimize the perplexity of a held-out data set.

In addition to the basic language models, we also developed a topic-adaptive language modeling technique using a multi-class *support vector machines* (SVM) -based topic classifier. The topics are organized as a manually constructed tree with 98 leaf nodes. To train the classifier, more than 20,000 Chinese news articles covering a wide range of topics are collected and annotated. The raw feature representing each training sample is a vector of terms given by our Mandarin text segmenter. An SVM is then trained to map from these feature vectors to topics. To reduce nuisance fea-

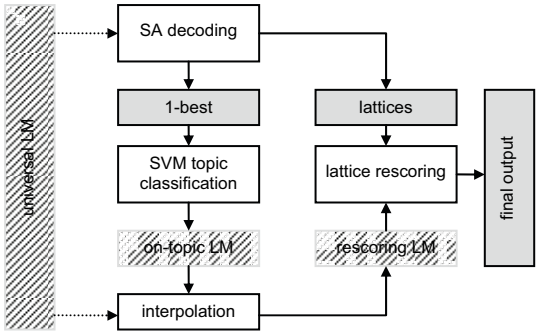

Fig. 2. Topic adaptation is carried out through lattice rescoring with an LM interpolated from the universal LM and a topic-specific LM. Topic classification is based on the 1-best word hypothesis given by the SA decoding output.

tures, words occurring in less than three documents are omitted. The overall classification accuracy of the topic classifier as measured by the F_1 measure is 0.8. An on-topic LM is trained for each of the 98 classes.

In decoding, the basic universal LM is first used to generate a word lattice and the 1-best hypothesis. The 1-best hypothesis is subsequently used for topic classification. Note that the change of topic occurs frequently in broadcast materials. Therefore, the classification is performed at the utterance level. Base on the classification result, an on-topic LM is selected from the 98 pre-trained LMs and interpolated with the universal LM. The resulting LM is used to rescore the lattices generated earlier to give the final recognition output. The process is shown in Fig 2.

3. SYSTEM BUILDING

3.1. Training Data

The acoustic modeling data is summarized in Table 1.

Table 1. Acoustic modeling data (with full transcripts).

Corpora	BN (Hours)	BC (Hours)
LDC1998T24	30.0	--
LDC2005E63	--	25.0
LDC2006E23	74.9	72.7
LDC2005S11	62.8	--
LDC2005E82	50.2	7.58
LDC2006E33	136.6	76.1
SATELLITE	50.0	--

A relatively small amount consists of broadcasts of news shows transcribed internally at IBM, and labeled "Satellite Data" below. From the data sources listed 550 hours were used to train our acoustic models, based on data that aligned to the transcripts using a set of boot models.

Our language model was built from all the acoustic transcripts, and additional text data that were used solely for language modeling purposes. This data is listed in Table 2.

Table 2. Language modeling Data.

Copora	Type	Number of words
LDC1995T13	Newswire	116M
LDC2000T52	Newswire	10.1M
LDC2003E03	News	1.4M
LDC2004E41	Newswire	17.1M
LDC2005T14	Newswire	245M
LDC2001T52	BN	4.7M
LDC2001T58	BN	3.1M
LDC2005E82 LDC2006E33	Blog & News-group	17.2M
SRI Web 20060522	Web	183M (char)
SRI Web 20060608	Web	5M (char)

3.2. System Description

The 16 KHz input signal is coded using 13-dememsional PLP features with a 25ms window and 10ms frame-shift. Nine consecutive frames are spliced and projected to 40 dimensions using LDA. The SI acoustic model has 10K quinphone states modeled by 150K Gaussian densities. The SA model uses a larger tree with 15K states and 300K Gaussians.

In addition to fully transcribed data, the training corpora also contain broadcast recordings with only closed captioning text. To take advantage of these data, *lightly supervised training* is applied.

The method relies on an automatic way to select reliable segments from the available data. First, we use the closed caption to build a *biased* LM. Then, the *biased* LM in conjunction with the existing acoustic model is used to decode the corresponding audio. The decoded text is aligned with the closed caption, and a segment is discarded unless it satisfies the following two criteria: (a) the longest successful alignment is more than three words; and (b) the decoding output ends on a silence word. The surviving data are deemed reliable and used for acoustic model training. This method is similar to those presented in [9] and [10].

It is observed that for broadcast news (BN) content, 55% of the closed caption data are eventually used in training, whereas for broadcast conversations (BC), only 21% survived the filtering process. In total, lightly supervised training increases the training set by 143 hours.

Our language model consists of an interpolation of eleven distinct models built from subsets of the training data. A held-out set with 31K words (61% BC, 39% BN) is used to determine the interpolation weights. The resulting LM has 6.1M n-grams, and perplexities of 735, 536, and 980 on RT04, 2006E10, and devo5bcm respectively.

4. EXPERIMENTAL RESULTS

Three test sets are used to evaluate the Mandarin broadcast transcription system. The first is the evaluation set from the Rich Transcription'04 (RT04) evaluation's Mandarin broadcast news task. It contains 61 minutes of data drawn from four BN recordings. The second test set, denoted "dev05bcm", contains five episodes of three BC programs. The total duration of this set is 3.5

hours. A third, 4.5-hour BN set "2006E10" is included to give more robust coverage of the BN content. The 2006E10 test set may be downloaded from the LDC, and includes RT04. The list of dev05bcm audio files was created at Cambridge University and distributed to GALE participants.

Recognition experiments on the three test sets are carried out following the pipeline shown in Fig. 1 in section 2. At the SA level, decoding using the ML acoustic model is done at after VTLN, after fMLLR, and after fMLLR to further understand the effect of each adaptation step on the Mandarin broadcast speech transpiration task. Except for VTLN decoding, the experiments are repeated using the MPE trained models and features. The recognition results are summarised in Table 4.

Table 3. Character error rates observed on the three test sets at different level of acoustic model refinement. The results indicate that discriminative training gives significant improvements in recognition performance.

System Build Level		RT04	dev05bcm	2006E10
SI:	--	19.4	28.3	19.6
	VTLN	18.0	26.7	18.1
	+fMLLR	16.5	24.9	17.1
SA:	+MLLR	15.7	24.3	16.7
	+fMPE+MPE+fMLLR	14.3	22.0	14.0
	+MLLR	13.7	21.3	13.8

As expected, the results show that BC data (dev05bcm) pose a greater challenge than the two BN sets. The results clearly confirm the effectiveness of the adaptive and discriminative acoustic modeling pipeline in the system. Furthermore, the overall trend of the CER as observed in each column is consistent across all three sets. In particular, we note that the MPE/fMPE algorithm gives a relatively large improvement to recognition performance on top of speaker adaptation. For instance, on 2006E10, discriminative training further reduces the CER by 2.9% absolute to 13.8% from the best ML models. Similarly, a 3.0% absolute reduction is achieved on the dev05bcm set. As a comparison, the MPE/fMPE gain observed in our Arabic broadcast transcription system is 2.1% absolute on the RT04 Arabic set.

To track the progress made in the GALE engagement, we compare the performance of the current system (06/2006) with our system at the end of 2005 (12/2005). The results are shown in Table 5. On RT04, a relative reduction in CER of 26.3% is observed. For reference, the best published numbers in the community on RT04 and dev05bcm are also listed [11].

Table 4. Comparing character error rates of the current system with the previous best-performing system and the best published results on the same test sets [11].

	System ID	RT04	dev05bcm	2006E10
SI:	12/2005	22.5	39.6	22.8
	06/2006	19.4	28.3	19.6
SA:	12/2005	18.6	34.5	20.0
	06/2006	13.7	21.3	13.8
	Best published	14.7	25.2	--

Finally, the topic-adaptive language modelling technique is evaluated by rescoring the SA lattices with (1) the LM that is topic-adapted to a given test utterance, and (2) an LM interpolated from the universal LM and a fixed set of eight topic-dependent LMs. On RT04, results show that the adaptive approach gives 0.4% absolute reduction in CER comparing with the non-adaptive counterpart.

5. CONCLUSIONS

In this work, we consider the Mandarin broadcast speech transcription task in the context of the DARPA GALE project. A state-of-the-art Mandarin speech recognition system is presented and validated on both BN and BC data. Experiments demonstrate that the MPE-based discriminative training leads to significant reduction in CER for this task. We also describe in this paper a topic-adaptive language modeling technique, and successfully apply the technique in the broadcast transcription domain.

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*, accepted for publication.

[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] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc ICASSP'96*, vol. 1, pp. 339-343, May 1996.

[4] 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.

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

[6] 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.

[7] S. F. Chen and J. T. Goodman, "An empirical study of smoothing techniques for language modeling," *Technical Report TR-10-98*, Computer Science Group, Harvard University, 1998.

[8] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," in *Proc. Eurospeech '97*, September 1997.

[9] L. Chen, L. Lamel, and J. L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Proc. ICASSP'04*, vol. 1, pp. 189-192, May 2004.

[10] 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.

[11] 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.

[12] B. Xiang, L.Nguyen, X. Guo, and D. Xu, "The BBN mandarin broadcast news transcription system," in *Proc. Interspeech'05*, pp. 1649-1652, September 2te.

[13] 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.