THE IBM 2004 CONVERSATIONAL TELEPHONY SYSTEM FOR RICH TRANSCRIPTION

Hagen Soltau, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon and Geoffrey Zweig

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

This paper describes the technical advances in IBM's conversational telephony submission to the DARPA-sponsored 2004 Rich Transcription evaluation (RT-04). These advances include a system architecture based on cross-adaptation; a new form of feature-based MPE training; the use of a full-scale discriminatively trained full covariance gaussian system; the use of septaphone cross-word acoustic context in static decoding graphs; and the incorporation of 2100 hours of training data in every system component. These advances reduced the error rate by approximately 21% relative, on the 2003 test set, over the best-performing system in last year's evaluation, and produced the best results on the RT-04 current and progress CTS data.

1. INTRODUCTION

One of the goals of the DARPA EARS program is to reduce the word error rate of transcribing telephone conversations to below 10%, and in support of this goal, NIST conducts periodic evaluations. This paper describes the IBM recognition system that was submitted to the 2004 evaluation, with special emphasis on newly developed techniques. Although the primary focus condition was on systems running in 20 times real-time, this paper focuses on IBM's 10xRT submission, which is somewhat more streamlined, and essentially as accurate.

The key design characteristics of the system include the following:

- A novel feature-space transform, termed fMPE, that is trained to maximize the MPE objective function. The fMPE transform operates by projecting from a very high-dimensional, sparse feature space derived from Gaussian posteriors to the normal recognition feature space and and adding the projected posteriors to the standard features. A system that uses fMPE+MPE training is better than a system using MPE alone by approximately 1.6% absolute, measured on the RT-03 test set.
- The use of a large, discriminatively trained full-covariance system having 143K 39-dimensional models. Using our fast decoding framework, decoding with this system requires 3.32 xRT.
- Static decoding graphs that use septaphone context, with both left and right cross-word context.
- System combination through cross-adaptation instead of acoustic rescoring of lattices. A cascade of speaker-adapted systems is used, with the output of one system being used to estimate the speaker-adaptive transforms for the next. This cascade consists of:
  1. a 150K Gaussian quinphone PLP speaker-independent system trained with MPE (all other systems are speaker adaptive);
  2. a 143K full-covariance Gaussian quinphone system built on fMPE features and trained with implicit-lattice MMI; and
  3. a 849K Gaussian septaphone PLP system trained with fMPE followed by traditional model-space MPE.
- Training of all system components on all available Fisher data.

Work funded by DARPA grant NBCH203001-2

2. A CROSS-ADAPTATION ARCHITECTURE

The 2004 IBM Rich Transcription system is organized around system combination through cross-adaptation. In common with typical evaluation systems [1, 2, 3], several different recognition systems are used in combination to produce the final output. Whereas typically this is done by generating lattices with one system and rescoring them with other systems, all communication in the 2004 IBM architecture is done through cross-adaptation. The sequence of acoustic models and decoding steps is described below.

The following acoustic models are used in the recognition process:

1. **SI.DC.PLP**: A speaker-independent model having 150K 40-dim diagonal-covariance mixture components and 8.0K quinphone context-dependent states, trained with MPE. Recognition features are derived from an LDA+MLLT projection from 9 frames of spliced, speaker-independent PLP features with blind cepstral mean normalization.
2. **SA.FC.fMPE**: A speaker-adaptive model having 143K 39-dim full-covariance mixture components and 7.5K quinphone context-dependent states, trained with MMI and fMLLR-SAT. Recognition features are derived from fMPE on an LDA+MLLT projection from 9 frames of spliced, VTLN PLP features with speech-based cepstral mean and variance normalization.
3. **SA.DC.fMPE+MPE**: A speaker-adaptive model having 849K 39-dim diagonal-covariance mixture components and 22K septaphone context-dependent states, trained with both fMPE and MPE, and fMLLR-SAT. Recognition features are derived from fMPE on an LDA+MLLT projection from 9 frames of spliced, VTLN PLP features with speech-based cepstral mean and variance normalization.

Processing of the speech then employs the following steps:

1. Segmentation of the audio into speech and non-speech.

2. Decoding with the SI.DC.PLP model.
3. Speaker adaptation and decoding with the SA.FC.fMPE model:
   (a) Estimation of speech-based cepstral mean and variance normalization and VTLN warping factors using the hypotheses from (2).
   (b) Estimation of fMPE, fMLLR and MLLR transforms for the SA.FC.fMPE model using the hypotheses from (2).
   (c) Decoding with the SA.FC.fMPE model.
4. Reestimation of speaker adaptive transforms and decoding with the SA.DC.fMPE+mMPE model:
   (a) Estimation of MLLR transforms using the features from (3b) and the hypotheses from (3c).
   (b) Lattice generation with the SA.DC.fMPE+MPE model.
5. Lattice rescoring with the LM described in Section 5.2.
6. Confusion network generation and the extraction of the consensus path.

The effect of cross-adaptation was studied on a combination of diagonal and full covariance models (table 1). Adapting the DC models on the errorful transcripts of the FC system led to a gain of 0.4% compared with self adaptation.

models/transcripts	FC	DC
FC	21.9	21.2
DC	21.0	21.4

Table 1. Error rates on RT-03. Comparison between Self- and Cross-Adaptation.

3. FMPE

The IBM evaluation system employs three forms of discriminative training. This section first mentions two traditional forms of discriminative training, and then presents a novel feature-based technique that is described fully in a companion paper [4].

The first form of discriminative training is implicit-lattice MMI, in which the denominator counts are collected by running a pruned forward-backward pass over a statically compiled decoding graph [5]. While each iteration of implicit-lattice MMI takes longer than a comparable pass of lattice-based MMI, the disk requirements of the implicit lattice technique are much smaller, which is advantageous when working with a large training set [6].

The second form of traditional discriminative training is MPE [7, 8]. This process used a lattice-based framework; in our implementation, lattices with fixed state alignments were used. Novel features include training with a pruned bigram language model with about 150K bigrams, instead of a unigram language model. The statistics in the MPE training were averaged over four sets of acoustic and language model weights, with the acoustic weight being either 0.10 or 0.16 and the language model weight being either 1.0 or 1.6. Experiments on a different data set indicated that this averaging may help when bigram lattices are used. In I-smoothing, the MMI rather than ML estimates of parameters were used for backing off. Also, during the update phase all variances were floored to the 20th percentile of the distribution of all variances in the appropriate dimension.

In addition to these traditional forms of discriminative training, the 2004 IBM system introduced a novel form of discriminative modeling, fMPE. This is a global discriminatively trained feature projection which works by projecting very high dimensional features based on posteriors of Gaussians down to the normal recognition feature dimension, and adding them to the normal features.

The technique is fully described in a companion paper [4], but briefly the process is as follows. A high dimensional feature vector is formed by evaluating 100,000 gaussians that broadly cover the input space. The likelihoods are normalized to sum to one, and the vector is expanded to 700,000 dimensions by appending averages of the posteriors of preceding and following frames. The feature projection, which is a matrix of size 700k x 39, is then trained by a form of gradient descent so as to optimize the MPE objective function, starting from a zero matrix (note that the features are added to the LDA+MLLT features so this is a reasonable initialization).

The method of fMPE may be compared with past work using neural-net posteriors as feature vectors [9]. However, previous methods either use only transformed posteriors as features, or concatenate posterior-derived features and standard recognition features. The new method differs in its high dimensionality and use of discriminative training to obtain a projection matrix.

ML Training	MPE	fMPE	fMPE+MPE
22.1	20.6	20.2	19.2

Table 2. Error rates on the 2003 EARS evaluation set.

Table 2 shows that fMPE reduces the word error rate by approximately 1.4% over the use of traditional MPE alone. Larger improvements can be obtained by using multiple layers of fMPE transform.

4. FULL COVARIANCE MODELING

One of the distinctive elements of IBM's 2004 system is the use of a full-scale acoustic model based on full-covariance Gaussians. Specifically, because of the availability of 2100 hours of training data (5), it was possible to build an acoustic model with 143,000 39-dimensional full-covariance mixture components. We have found that full covariance systems are slightly better than similarly sized diagonal covariance systems, and in addition are beneficial for cross-adaptation. To construct and use this model, the following problems were solved.

- Speed of Gaussian evaluation. Firstly, we based the computations on a Cholesky decomposition of the inverse covariance matrix. This allows pruning the likelihood computation for a mixture component as soon as the partial sum across dimensions falls below a threshold. Secondly, we used a hierarchical Gaussian evaluation setup described in [10]. By combining these two approaches, the runtime for full decoding was brought to 3.3 times real-time without loss in accuracy.
- Discriminative training. Using the speedup just mentioned with tight beams and a small decoding graph, implicit lattice MMI [5] was easily possible. The following MMI update equations were used for the means and the covariance matrices:

$$\hat{\mu}_i = \frac{\theta_i^{num}(\mathbf{x}) - \theta_i^{den}(\mathbf{x}) + D\mu_i}{\theta_i^{num} - \theta_i^{den} + D} \qquad (1)$$

$$\hat{\mathbf{\Sigma}}_i = \frac{\theta_i^{num}(\mathbf{x}\mathbf{x}') - \theta_i^{den}(\mathbf{x}\mathbf{x}') + D(\mu_i\mu_i' + \mathbf{\Sigma}_i)}{\theta_i^{num} - \theta_i^{den} + D} - \hat{\mu}_i\hat{\mu}_i' \quad (2)$$

where the θ's represent the mean, variance and prior statistics for the numerator and the denominator. D is chosen on a per Gaussian basis such as to ensure that $\hat{\mathbf{\Sigma}}_i$ is positive definite and that its minimum eigenvalue is greater than a predefined threshold. In adddition, I-smoothing was used for the numerator counts [7]. The effect of MMI is illustrated in table 3 for both standard SAT and SAT-fMPE features.

	ML Training	MMI
SAT	23.2	22.1
SAT-fMPE	21.4	20.0

Table 3. Error rates on the 2003 EARS evaluation set.

- MLLR transform estimation. Only the on-diagonal elements of the covariance matrix were used to estimate MLLR transforms; this produced WER reductions of approximately 1% absolute, in line with expectations.

5. TRAINING DATA AND SYSTEM BASICS

5.1. Training Data

5.1.1. Acoustic Model Data

The acoustic training set used data from 5 sources: Fisher parts 1-7, Switchboard-1, BBN/CTRAN Switchboard-2, Switchboard Cellular, and Callhome English.

The Fisher transcripts were normalized using a collection of 840 rewrite rules, some of which corrected classes of errors. 41 conversation sides in the original collection were rejected because they had insufficient quantities of data (less than 20 s. of audio), and an additional 47 hours of data containing words occurring 4 times or less in the whole corpus were rejected.

We used ISIP's 25 October 2001 release of Switchboard transcripts for the Switchboard-1 data, with a few manual corrections of transcription errors.

The BBN/CTRAN Switchboard-2 transcripts and LDC transcripts for Switchboard Cellular and Callhome English were normalized to follow internal conventions, and a few manual corrections were made.

In addition, the full collection of audio data was resegmented such that all training utterances had nominally 15 frames of silence at the beginning and end, and all single-word utterances were discarded [11]. Following normalization, roughly 2100 hours of training data remain.

5.1.2. Language Model Data

We used seven data sources for our language model training:

1. SWB - LDC transcripts of Switchboard-1, Switchboard Cellular and Callhome English.
2. BBN - BBN/CTRAN transcripts of Switchboard-2.
3. BN - Broadcast News transcripts.
4. FSH - Fisher parts 1-7.
5. UW191 - 191M words of 'Switchboard-like' web data collected by the University of Washington.
6. UW175 - an older collection of 175M words of 'Fisher-like' web data collected by the University of Washington.
7. UW525 - a newer collection of 525M words of 'Fisher-like' web data collected by the University of Washington.

5.2. System Basics

We use a recognition lexicon of 30.5K words which was generated by extending our RT-03 lexicon to cover the 5000 most frequent words in the Fisher data. The lexicon contains a total of 33K variants (1.08 variants per word). Pronunciations are primarily derived from PRONLEX, with the manual addition of a few variants to cover reduced pronunciations that are common in conversational American English. Pronunciation variants have weights based on their unigram counts in a forced alignment of the acoustic training data.

5.2.1. Acoustic Modeling

The raw acoustic features used for segmentation and recognition are perceptual linear prediction (PLP) features as described in [10]. No echo cancellation was performed.

The features used by the speaker-independent system are mean-normalized on a conversation side basis. The features used by the speaker-adapted systems are mean- and variance-normalized on a conversation side basis, but normalization statistics are accumulated only for frames labeled as speech in the speaker-independent pass.

Words are represented using an alphabet of 45 phones. Phones are represented as three-state, left-to-right HMMs. With the exception of silence and noise states, the HMM states are context-dependent, and may be conditioned on either quinphone or septaphone context. In all cases, the phonetic context covers both past and future words. The context-dependent HMM states are clustered into equivalence classes using decision networks.

Context-dependent states are modeled using mixtures of either diagonal-covariance or full-covariance Gaussians. For the diagonal-covariance systems, mixture components are allocated according to a simple power law, $m = min(M, ceil(k * N^{0.2}))$, where m is the number of mixture components allocated to a state, M is the maximum number of mixtures allocated to any state, N is the number of frames of data that align to a state in the training set, and k is a constant that is selected to set the overall number of mixtures in the acoustic model. Initial maximum-likelihood training of the diagonal-covariance systems is based on a fixed, forced alignment of the training data at the state level [11], and uses an iterative mixture-splitting method to grow the acoustic model from a single component per state to the full size. Typically, maximum-likelihood training concludes with one or two passes of Viterbi training on word graphs. All training passes are performed over the full 2100-hour acoustic training set.

We use two forms of feature-space normalization, vocal tract length normalization (VTLN) [12] and feature-space MLLR (fMLLR, also known as constrained MLLR) [13], in the context of speaker-adaptive training to produce canonical acoustic models in which some of the non-linguistic sources of speech variability have been reduced.

The VTLN warping is implemented by composing the 21 piecewise linear warping functions with the Mel filterbank to generate 21 different filterbanks. The warping function is chosen to maximize the likelihood of frames that align to speech under a

model that uses a single, full-covariance Gaussian per context-dependent state to represent the class-conditional distributions of the static features. Approximate Jacobian compensation of the likelihoods is performed by adding the log determinant of the sum of the outer product of the warped cepstra to the average frame log-likelihood.

We do a single pass of MLLR adaptation for each conversation side, using a regression tree to generate transforms for different sets of mixture components. The regression tree is an 8-level binary tree that is grown by pooling all of the mixture component means at the root node, then successively splitting the means at each node into two classes using a soft form of the k-means algorithm. The MLLR statistics are collected at the leaves of the tree and propagated up the tree until a minimum occupancy of 3500 is obtained and a transform is generated.

5.2.2. Language Modeling

The IBM 2004 system uses two language models: a 4.1M n-gram language model used for constructing static decoding graphs, and a 100M n-gram language model that is used for lattice rescoring. Both language models are interpolated back-off 4-gram models smoothed using modified Kneser-Ney smoothing. The interpolation weights are chosen to optimize perplexity on a held-out set of 500K words from the Fisher corpus. The interpolation weights of the decoding graph and rescoring language models are given in Tables 4.

	LM1	LM2
SWB	0.10	0.07
BBN	0.15	0.05
BN	0.05	0.04
FSH	0.55	0.71
UW191	0.15	-
UW175	-	0.02
UW525	-	0.11

Table 4. Interpolation weights for the decoding graph LM (LM1) and rescoring LM (LM2).

6. FULL SYSTEM RESULTS

Word error rates at the different system stages are presented in Table 5 for the 2003 test set provided by NIST, and the 2004 development set. Numbers are given at the different stages: SI is speaker-independent decoding; FC is the full-covariance system; DC is the diagonal covariance fMPE + mMPE system and LN+CN denotes LM rescoring followed by confusion network generation.

	RT03	RT03-FH	DEV04	RT-04
SI	28.0	23.2	23.5	26.7
FC	19.7	16.0	16.6	18.8
DC	17.4	14.1	14.5	16.4
LM + CN	16.1	12.4	13.0	15.2

Table 5. Error rates at different system stages.

7. CONCLUSION

This paper has described a simple and effective recognition system for conversational telephony. Notable features include fMPE, full-covariance gaussians, the use of septaphone acoustic context, and system combination through cross-adaptation. Error rates below 15% are reported for EARS conversational telephony test sets.

8. REFERENCES

[1] A. Stolcke, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, and J. Zheng, "Speech to text research at SRI-ICSI-UW," in *NIST RT-03 Workshop*. May 2003, DARPA.

[2] S. Matsoukas, R. Iyer, O. Kimball, J. Ma, T. Colhurst, R. Pasad, and C. Kao, "BBN CTS english system," in *NIST RT-03 Workshop*. May 2003, DARPA.

[3] P. Woodland, G. Evermann, M. Gales, T. Hain, R. Chan, B. Jia, DY. Kim, A. Liu, D. Mrva, D. Povey, KC Sim, M. Tomalin, S. Tramter, L. Wang, and K. Yu, "CU-HTK STT systems for rt03," in *NIST RT-03 Workshop*. May 2003, DARPA.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *ICASSP-2005*. 2004, Submitted.

[5] J. Huang, B. Kingsbury, L. Mangu, G. Saon, R. Sarikaya, and G. Zweig, "Improvements to the IBM hub5e system," in *NIST RT-02 Workshop*. 2002, DARPA.

[6] B. Kingsbury, S. Chen, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Training a 2300-hour fisher system," in *EARS STT Workshop*, 2004.

[7] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *ICASSP-2002*, 2002.

[8] D. Povey, *Discriminative Training for Large Voculabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[9] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP-2000*, 2000.

[10] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Eurospeech-2003*, 2003.

[11] H. Soltau, H. Yu, F. Metze, C. Fuegen, Q. Jin, and SC. Jou, "The 2003 ISL rich transcription system for conversational telephony speech," in *ICASSP-2004*, 2003.

[12] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *ICASSP-1996*, 1996.

[13] M.J.F. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," in *Tech. Report CUED/F-INFENG/TR291*. 1997, Cambridge University.