TRANSCRIPTION OF BROADCAST NEWS - SOME RECENT IMPROVEMENTS TO IBM'S LVCSR SYSTEM

L. Polymenakos, P. Olsen, D. Kanvesky, R. A. Gopinath, P. S. Gopalakrishnan and S. Chen

IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email:rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

This paper describes extensions and improvements to IBM's large vocabulary continuous speech recognition (LVCSR) system for transcription of broadcast news. The recognizer uses an additional 35 hours of training data over the one used in the 1996 Hub4 evaluation [7]. It includes a number of new features: optimal feature space for acoustic modeling (in training and/or testing), filler-word modeling, Bayesian Information Criterion (BIC) based segment clustering, an improved implementation of iterative MLLR and 4-gram language models. Results using the 1996 DARPA Hub4 evaluation data set are presented.

1. INTRODUCTION

Recently interest in large vocabulary continuous speech recognition recognition (LVCSR) research has shifted from read speech data to speech data found in the real world - like broadcast news (BN) over radio and TV and conversational speech over the telephone. Considerable amount of both acoustic (approximately 100 hours of which about 70% is usable) and linguistic (approximately 400 million words) training data for BN has been made by the Linguistic Data Consortium (LDC) in the context of DARPA sponsored Hub4 evaluations of LVCSR systems on BN [1]. As has been studied and reported by several researchers [4, 7, 11, 10, 8, 9], BN transcription poses several challenges to LVCSR systems. The speech data exhibits a wide variety of speaking styles, environmental and background noise conditions and channel conditions. The general approach has been to classify the BN data into a set of homogeneous conditions and to build acoustic models (AMs) for each condition. Test data is then segmented and classified along conditions and an appropriate acoustic model used for each condition. One particular classification scheme for BN news data that has been used in the DARPA sponsored Hub4 BN evaluation in 1996 splits the speech data along the so-called F-conditions [1]: prepared speech (F0), spontaneous speech (F1), low fidelity speech, including telephone channel speech (F2), speech in the presence of background music (F3), speech in the presence of background noise (F4), speech from non-native speakers (F5) and FX - all other speech. The 1996 Hub4 Unpartitioned Evaluation (UE) and Partitioned Evaluation (PE) test data set forms a standard test set for evaluating LVCSR systems. The only difference between the UE and PE tests is that in the latter, the data is segmented and classified into F-conditions manually, while in the former this has to be done automatically if necessary. A comparison gives information on how well automatic segmentation schemes work for BN news transcription.

For the UE test, in the past we have used a two-stage approach [4]. The speech data is first segmented into high bandwidth speech (clean), low bandwidth speech (telephone), and music. The music segments are removed and the high and low bandwidth speech segments are then decoded using models trained (or adapted) on high and low bandwidth speech respectively. This is because of the inadequacy of current segmentation algorithms to separate out other F-conditions; it is relatively easy to detect music or telephone channel speech.

For the PE test, in the past we have built condition specific models for each condition using MAP and MLLR. This is because there is not sufficient training data to independently build models for each F-condition; besides, it may not be the best way to handle the problem.

Our current approach for both the UE and PE tests is to use a single robust model built on all the available training data. Speaker/condition-adapted (SAT) training [6], while appropriate for this purpose, is not used in the model described in this paper. For both the PE and UE tests, iterative MLLR is used to adapt the baseline robust model for both the speaker and the F-condition. For the UE test, the data is still, however, segmented into low-bandwidth and high-bandwidth segments. The segments are then clustered into homogeneous groups (the same speaker or environmental condition) before iterative MLLR is applied.

In this paper we present algorithmic improvements to the baseline model used in the UE/PE evaluations in 1997. Some of the improvements are: use of optimal feature spaces for modeling gaussian distributions (in a maximum likelihood sense), filler-word modeling, use of Bayesian Information Criterion for segment clustering, and an improved implementation of iterative MLLR. The focus of the research effort has been on improving baseline recognition accuracy for clean speech (i.e., the F0 and F1 conditions).

2. OVERVIEW OF THE LVCSR SYSTEM

The IBM LVCSR system uses acoustic models for subphonetic units with context-dependent tying [2, 3] for details). The instances of context dependent sub-phone classes are identified by growing a decision tree from the available training data [2] and specifying the terminal nodes of the tree

Acoustic Model	F0	F1
AM-base	21.4	30.3
AM-0(4.0K)	21.3	29.7
AM-1(2.0K)	22.6	31.0
AM-2(3.5K)	21.1	29.1
AM-3(7.3K)	21.9	30.3

Table 1. Comparison of Decision Tree Sizes: AM-base - trained on WSJ, AM-0 - trained on BN with 4K leaves, AM-1 trained on F0+F1 portion with 2K leaves, AM-2 same as AM-1 with 3.5K leaves, AM-3 same as AM-2 with 7.3K leaves.

as the relevant instances of these classes. The acoustic feature vectors that characterize the training data at the leaves are modeled by a mixture of Gaussian pdf's, with diagonal covariance matrices. The HMM used to model each leaf is a simple 1-state model, with a self-loop and a forward transition.

The recognizer used in the 1996 evaluation had 5.7K HMM states (or leaves) and 170K gaussians. The decision tree for the HMM states was built WSJ0+1 data. The gaussian mixtures, however were trained on the approximately 35 hours of BN training data distributed by LDC in 1996. For the PE test, the models were further adapted to each focus condition and for the UE test to high/low bandwidth speech using a combination of MAP and MLLR [15, 14, 7] adaptation.

3. ACOUSTIC MODELING

A new baseline acoustic model (AM-base) with 90K gaussians was built using all the 70 hours of training data (including the 35 hours of additional data distributed in 1997) by rebuilding gaussian mixtures for the 5741 HMM states. Since these states were constructed from WSJ data we built two new decision trees for context clustering, one based on just the clean (F0+F1) training data and the other based on all the training data. Gaussian mixtures were then estimated using the EM algorithm and the performance for various model sizes were evaluated. Experimental results for the F0 and F1 focus conditions on the PE test are shown in Table 1. The language model (LM) used in these experiments is LM-base (see below) and there are about 90K gaussians in each of the acoustic models. Firstly, notice that building the decision tree with the BN data improves error rate on both F0 and F1 (WER with AM-base is worse than WER with AM-0 or AM-2). The improvements are more on F1 (spontaneous speech) because of the new realizations of context-dependent sub-phonetic units vis à vis WSJ training data. Secondly, not using the training data for the other F-conditions in tree building gives more gain (AM-0 vs. AM-2). This is probably because some of the HMM states are now modeling realizations of phones in specific environmental conditions. The best results were obtained with a system with about 3.5K HMM states (AM-2).

3.1. Filler Models

The training data is transcribed with breath and filled-pauses allowing us to build models for filler words. Filler words are transcribed using our usual phone set of 51 phones in the dictionary. To the decision trees that take sub-phonetic units

Acoustic Model	F0	F1
AM-base	21.4	30.3
AM-4	21.0	29.0
AM-2	21.1	29.1
AM-5	21.0	28.9

Table 2. % word error rate with filler word models: AM-base and AM-2 do not use filler models. AM-4 is AM-base with filler models and AM-5 is AM-2 with filler models.

to the HMM states, new states were added for each occurrence of a phone within a particular filler word. The models for these states were initialized by those of some other state of the same sub-phonetic unit. Standard Baum-Welch reestimation is then used to estimate the models. Filler models seem to improve the performance on spontaneous speech without degrading the performance on prepared speech when the base models was AM-base. However, the gain is marginal when the base models used is AM-2. This is presumably because AM-base HMM states were built on WSJ data while AM-2 HMM states were built on the BN training data and hence some states were already modeling filler words. Results are summarized in Table 2.

3.2. Optimal Features Spaces for Modeling

The number of gaussians used in current LVCSR systems implies (from data insufficiency, storage and computational considerations) that only diagonal gaussian models can be used. With full-covariance gaussian models linear transformations of the feature space clearly does not lead to a better model. In fact, if the transformation is unimodular (or volume-preserving) the likelihood is exactly the same in all transformed spaces. However, with diagonal gaussian models one can ask among all possible transformed feature spaces which is the one where the diagonal assumption is "most valid". If the transformation is unimodular (required only to simplify the argument), then, in each transformed space there is a loss in likelihood with respect to full-covariance modeling (which is a constant). One can therefore find a transformed space in which the loss in likelihood is least (for details see [12]). This gives a single global transformation on the feature space. Notice however, the gaussians can be clustered into groups and each group can be modeled in its own feature space. Since there is more flexibility in this case the loss in likelihood is less. In the extreme case where each gaussian has its own feature space transformation one can choose the transformation to be projection onto the eigenbasis of its covariance matrix and the likelihood of the data is the same as full-covariance likelihood. However, from computational and storage points of view this is exactly as expensive as full-covariance modeling. If (x_i, l_i) is the labeled (at HMM state level) training data, $i \in \{1, 2, \ldots, N\}, x_i \in \mathbb{R}^d, l_i \in \{1, 2, \ldots, J\}$, and $c_j \in \{1, 2, \ldots, K\}$ is the class cluster (or transformation id) map, and Σ_j is the covariance at state j (we are assuming a single gaussian at each state for simplicity), then the likeli-

Acoustic Model	F0	F1
AM-2(baseline)	21.1	29.1
AM-6(1 transform)	19.3	28.4
AM-7(4 transforms)	19.4	29.0

Table 3. Optimal Feature Spaces for HMM state clusters: a) AM-2 - baseline b) AM-6 - single transform c) AM-7 - 4 transforms

Acoustic Model	F0	F1
AM-6	19.3	28.4
AM-8	19.3	27.9

Table 4. Supervised adaptation on F0+F1 using MLLR (% WER): AM-6 - baseline, AM-8 - adapted models.

hood of the training data is given by [12]:

$$p^{\star}_{diag}(x_1^N) = g(N,d)\prod_{j=1}^{J}\left|A_{c_j}\right|^{N_j}\left|diag(A_{c_j}\bar{\Sigma}_j A_{c_j}^T)\right|^{-\frac{N_j}{2}}.$$

Maximizing the above expression numerically gives the optimal choice of transforms A_k, $k \in \{1,2,\ldots,K\}$. In our experiments, after the transform is obtained this way, using single-pass-retraining from a baseline system, gaussian mixture models are built for each HMM state using the new (state-dependent) feature space. Here we present results when using one (AM-6) and four transformations (AM-7) on the AM-2 baseline acoustic models (see Table 3). For the latter case one transform each was used for all the gaussians corresponding to a) stop-consonants and flaps, b) fricatives, c) vowels and dipthongs, and d) nasals, glides, and silence respectively. This notion of optimal-feature spaces is the same as the notion semi-tied full-variances [17].

3.3. Supervised adaptation F0 and F1

All of the acoustic models above were built on training data from all the F-conditions. Since we are especially interested in the performance of our LVCSR system on F0 and F1 the model AM-5 was further adapted using the F0 and F1 portion of the training data (about 60%). The performance of the baseline model (AM-6) and the adapted model (AM-8) are shown in Table 4.

4. SEGMENTATION AND CLUSTERING

4.1. Segmentation

Gaussian mixture models for low bandwidth speech, high bandwidth speech and pure music are used to segment the data [4]. Besides, the test data is decoded with a a small vocabulary (5K words) and a small acoustic model set to obtain silence segments. This information is used to prevent segment boundaries from splitting words. The UE and PE test data is identical (except for the side-information of the F-conditions in PE). Therefore a comparison of UE and PE performance gives and evaluation of the segmentation procedure. Experimental tests were conducted on the acoustic model AM-6 described earlier and the results are shown in Table 5 The segmentation procedure leads to a loss of about 1% in accuracy across all conditions.

Ac. Model	F0	F1
AM-6+true cluster	17.5	24.8
AM-6+auto cluster	17.5	24.6

Table 6. Manual vs Automatic Clustering Performance

4.2. Unsupervised Adaptation on Test Data

Adaptation schemes like MLLR [14] adapt the means and variances of the gaussian models using linear transformations. If there are too many adaptation parameters or too little adaptation data, then, the adaptation tends to learn the adaptation data transcriptions quickly. To alleviate this problem we can decrease the number of adaptation parameters or increase the amount of adaptation data. The former is accomplished in the context of an iterative MLLR scheme where there are $2^i + 1$ transforms at the i^{th} iteration for 2^i non-silence phonetic sub-units and one transformation all the phonetic sub-units of silence. The transformations are applied only to the means; the variances are just scaled to maximize the likelihood on the test data ([16]). Increase in the amount of adaptation data is accomplished by clustering together similar the segments using a Bayesian Information Criterion (BIC) [13].

4.3. Clustering for Unsupervised Adaptation

The segments are clustered using a standard maximum-linkage bottom-up-clustering procedure with a single gaussian model for each segment and log-likelihood ratio distance measure. The termination for this bottom-up-clustering procedure was determined to maximize the BIC criterion [13]. BIC is a likelihood criterion penalized by the model complexity (the number of parameters in the model). At each stage in the bottom-up-clustering process the increase in BIC value is computed and the process is terminated when this increase is negative. It can be easily be shown that the increase in BIC value by merging two clusters is given by

$$-n\log|\Sigma| + n_1\log|\Sigma_1| + n_2\log|\Sigma_2| + N(d + \frac{d(d+1)}{2}),$$

where $n = n_1 + n_2$ is sample size of the merged node, Σ is the covariance matrix of the merged node and N is the total number of samples from all the segments. This gives, in principle, a threshold-free approach to clustering.

To study the effectiveness of clustering, the F0 and F1 segments of PE test were clustered by hand (28 clusters) and by using the algorithm described above (31 clusters). The word error rate (WER) after iterative MLLR adaptation is nearly the same as seen in Table 6. In contrast the result of clustering all the PE segments automatically (79 clusters) is shown in Table 7 with single and multiple iterations of MLLR. For comparison the baseline numbers are also given.

5. LANGUAGE MODELING

The Language Model has a vocabulary of 65K most frequent words from the BN language model corpus distributed by LDC in 1996. The baseline language model (LM-base) is the

Test	Total	F0	F1	F2	F3	F4	F5	FX
PE	28.2	18.6	25.1	34.8	24.7	34.8	29.1	54.2
UE	29.5	19.4	26.0	39.0	27.2	36.2	24.1	55.2

Table 5. Segmentation Accuracy: PE vs. UE (% WER).

Ac.Model	Total	F0	F1	F2	F3	F4	F5	FX
AM-6+auto cluster	29.8	18.8	27.0	39.1	29.9	36.3	30.1	54.2
AM-6+auto cluster +MLLR	27.8	17.9	25.8	33.1	26.6	35.2	27.8	49.9
AM-6+auto cluster + iterative MLLR	27.0	17.3	24.9	32.5	26.5	35.7	26.1	47.5

Table 7. Clustering for Unsupervised Adaptation (% WER): AM-6-auto - baseline with clustering, AM-6-auto cluster +MLLR1 - additionally one iteration of MLLR, M-6-auto cluster +iterative MLLR - iterative MLLR.

Lang. Model	F0	F1
LM-base	21.0	29.1
LM-base+4g	20.8	28.7
LM-base+4g+ac	20.7	28.6

Table 8. Mixture LM with 4-gram and acoustic transcriptions

one used in the 1996 evaluation [4]. With the same training data a standard 4-gram deleted interpolation LM was built (LM-4g). This component was added to LM-base to create LM-base+4g. This LM was further mixed with a small LM built from the 70 hours of acoustic training data transcriptions (LM-base+4g+ac). Mixing the 4-gram LM and the acoustic transcriptions LM to the baseline LM gives minor improvements to the recognition performance as seen in Table 8. The acoustic model used in these experiments was AM-6.

6. CONCLUSION

Transcription of broadcast news poses several challenges. This paper presented improvements and extensions of the IBM LVCSR system used in the 1996 DARPA Hub4 evaluation. Optimal feature spaces for modeling is shown to lead to significant improvements in the baseline accuracy. However, further improvements are required, especially, in robustness to channel and noise degradations.

REFERENCES

[1] D. Pallet, "Overview of the 1997 DARPA Speech Recognition Workshop", Proc. of DARPA Speech Recognition Workshop, Feb 2-5, Chantilly VA, 1997.

[2] L. R. Bahl et al., "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. ICASSP, 1994.

[3] L. R. Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proc. ICASSP, pp 41-44, 1995.

[4] P. S. Gopalakrishnan et al., "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System," Proc. ARPA SLT Workshop, Feb 1996.

[5] P. S. Gopalakrishnan, et al., "Acoustic Models Used in the IBM System for the ARPA Hub 4 Task," Proc. ARPA SLT Workshop, Feb 1996.

[6] T. Anastasakos, et al., "A Compact Model for Speaker-Adaptive Training", Proc. ICSLP-96.

[7] R. Bakis et al., Transcription of BN Shows with the IBM LVCSR System", Proc. DARPA Sp. Reco. Workshop, 1997.

[8] F. Kubala et al., "The 1996 BBN Byblos Hub4 Transcription System", Proc. DARPA Sp. Reco. Workshop, 1997

[9] P. Placeway et al., "The 1996 Hub4 Sphinx System", Proc. DARPA Sp. Reco. Workshop, 1997.

[10] P. C. Woodland et al., "The Development of the 1996 HTK Broadcast News Transcription System", Proc. DARPA Sp. Reco. Workshop, 1997

[11] J. L. Gauvain et al., "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System", Proc. DARPA Speech Recognition Workshop, 1997.

[12] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions for Classification", submitted to ICASSP 1997.

[13] S. Chen et al, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", submitted to ICASSP 1997.

[14] C. J. Legetter et al., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.

[15] J. L. Gauvain et al., "Maximum-a-Posteriori estimation for multivariate Gaussian observations of Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp 291-298, Apr 1994.

[16] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Tech. Rep., CUED/FINFENG/TR291, Cambridge Univ., 1997.

[17] M. J. F. Gales, "Semi-tied Full-covariance matrices for hidden Markov Models", Tech. Rep., CUED/FINFENG/TR287, Cambridge Univ., 1997.