The AT&T LVCSR-2000 System

Andrej Ljolje
alj@research.att.com

Donald M. Hindle
hindle@research.att.com

Michael D. Riley
riley@research.att.com

Richard W. Sproat
rws@research.att.com

AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932-0971, USA

ABSTRACT

We describe the AT&T recognition system used in the Large Vocabulary Conversational Speech Recognition (LVCSR-2000) evaluation. It is based on multi-pass rescoring of weighted Finite State Machines (FSMs) using progressively more accurate models. The final stages, which provide highest recognition accuracy, use combined outputs of several models that were designed so that they are as different from each other as possible. Acoustic models used in the system varied along three different dimensions aimed at providing maximally different acoustic models: gender dependency, cepstral variance normalization and triphone vs. pentaphone. We used six out of the total of eight possible combinations of model features. The last steps use the outputs from all the models to improve the Maximum Likelihood Linear Regression (MLLR) adaptation step and finally the recognition results with the adapted models are all used to provide the single best combined recognition hypothesis.

1. INTRODUCTION

Large Vocabulary Conversational Speech Recognition (LVCSR) evaluations are the result of the long-term effort to improve the recognition accuracy of truly spontaneous conversational speech recorded over the telephone.

The training data and the evaluation data consist of parts of two databases. The *Switchboard* database is a collection of conversations between subjects on a predefined topic. The *Call Home* database is a collection of conversations between subjects and their family members/friends with no constraints on the topic. Although there is an order of magnitude more *Switchboard* training data, there is the same amount of evaluation data from the two data sets.

The combination of unconstrained topic and vocabulary, limited bandwidth, network distortions, environment noise and the spontaneity of speech (including all disfluencies that occur in spontaneous speech), result in relatively low recognition accuracy rates, despite the ability to use multi-pass strategies and allowing non-real-time scenarios for off-line recognition.

We report on many improvements to our system since the LVCSR-98 evaluation [4]. In particular, we describe the ways to combine recognition results obtained with differently trained acoustic models (both word lattices and one-best results) to improve recognition results relative to the best achieved result with any of the available models.

2. ACOUSTIC TRAINING DATA

The official training set is open to any publicly available speech and text data. However, in practice, most sites participating in the evaluations use the original collections of recordings from the *Switchboard* database (*Switchboard 1*). *Switchboard 2*, which are the more recent recordings, are reserved for the evaluation data. In addition, a predefined number of recorded *Call Home* conversations are included in the training data set.

The recordings were originally done in stereo, with each channel corresponding to one side of the telephone conversation, recording the complete conversation. In order to train acoustic models, lexical transcription and segmentation into *turns* is required, together with a dictionary for the words in the transcriptions.

In the past, we obtained suitable transcriptions and segmentation into turns form BBN, but this covered only some of the training data. For this evaluation we mostly used the segmentations and transcriptions from the effort at Mississippi State University (MSU) to clean-up the Switchboard transcriptions, as available in November, 1999. More specifically, we used all of the cleaned-up conversations provided by MSU, or we used the original BBN transcriptions, or if neither was available we used the transcriptions provided by MSU, but which had not been cleaned-up. We used the MSU version of the pronlex dictionary, augmented by the few words that appeared in the transcriptions, but were not part of the dictionary, resulting in the total of 45642 lexical items. We used 2245 conversation sides that were cleaned up at MSU, 1331 conversation sides as provided to us by BBN in 1997, and 999 conversation sides provided by MSU, but that were not cleaned-up. In addition we used 234 conversation sides of *Call Home* data. No attempt was made to improve the quality of the transcriptions or segmentations into turns in any way whatsoever. A few conversation sides were eventually discarded, as there was insufficient data to successfully perform adaptation, and thus could not be used in training involving adaptation.

3. LANGUAGE MODELS

We used two kinds of language models (LMs) in the evaluation. The first was a smaller and simpler 3-gram language model which was used in all of the recognition passes. The larger and more complex 6-gram language model (and the adapted version of the same) was only used to rescore existing word lattices which were generated using the 3-gram language model. This is achieved by taking the word lattices generated at a recognition/rescoring pass which have combined acoustic, duration and language model costs at every word arc, removing the 3-gram language model cost and replacing it by the 6-gram language

Additional LM Training Data			
Show	Source	No. words	Type
Soap Operas		4.5 Million	Closed Capt.
All My Children	http://members.tripod.com/~LauraAMC		
Another World	http://members.tripod.com/176 Laura_AW		
General Hospital	http://members.tripod.com/176 GHTranscripts		
One Life to Live	http://members.tripod.com/176 LauraOLTL		
Port Charles	http://members.tripod.com/176 PCTranscripts		
Friends	http://www.geocities.com/Hollywood/9151	141K	Transcripts
Sabrina the Teenage Witch	http://www.bccnet.force9.co.uk/transcripts/	243K	Transcripts
Real Hollywood Talk	http://www.universalstudios.com/unichat.30/newchat/transcripts	1.3 Million	Transcripts

Table 1: Sources for Additional training Data for Improving Language Models

model cost.

The language model used in the recognition/rescoring passes was a simple Katz backoff 3-gram model, based on training from three sources: *Switchboard* transcripts (3.5M words), *Call Home* transcripts (.3M words), and *Broadcast News* transcripts (165M words). The contributions of the training data from the three sources were effectively weighted in a ratio of (1:1:.15). The 3-gram model was shrunken using the technique of Seymore and Rosenfeld [1], giving a model with 497855 states and 1554689 arcs (45642 1-grams, 452212 2-grams and 558979 3-grams).

For training the larger, recoring language model, in addition to the training text used for the first pass model (from *Switchboard, Call Home and Broadcast News*), transcripts from various television shows were added, as shown in Table 1.

Using this additional data, a 6-gram backoff language model was created, with 9.2M states, 18.6M arcs, and a perplexity of 95 on a development test set (45642 1-grams, 1238109 2-grams, 2431313 3-grams, 774844 4-grams, 179545 5-grams and 28717 6-grams).

In the final pass, after rescoring lattices with this 6-gram language model, the LM scores were further adjusted. For each conversation, all the words in the best path transcriptions of all turns were collected. For each word, a weighting factor was assigned based on the likelihood that a word will appear more than once in a conversation. The word lattices were then rescored with each candidate word arc LM cost multiplied by this factor.

More specifically, we adopted a minimal approach to trigger word modeling, namely modeling the likelihood that a word is its own trigger (the likelihood that having seen a word once in a conversation that same word will be seen again). To model this effect we add to each word in the 6-gram rescored word lattice a word-boosting factor, based on the number of conversations in which that the word appears more than once in *Switchboard* and *Call Home* training transcriptions. The boosting factor is estimated as FW2 / FC2, where FW2 = the count of the word occurring more than once in a conversation in the training, and FC2 = the number of conversations that the word appears more than once in the training. This boosting weight is normalized by the overall word frequency and the number of conversations (since this is already accounted for in the language model). To use this weighting, for each conversation, the the set of words in the best path for all the segments is determined. Then for each segment, each word in the word lattice is reweighted by the boosting weights for this word set; words not in the best path are not reweighted. Note, this word-boosting does not preserve the probability interpretation of the language model, and indeed, the boosting was scaled by a factor selected based on explorations with various language models for the development test set. The improvement in word error rate from this word boosting on the development test set is small (0.0-0.2%) and doesn't seem entirely stable.

Language Model	Word Error Rate (%)
3-gram	44.0
6-gram	42.9
6-gram + TV shows	42.6
6-gram + TV shows + adaptation	42.5

Table 2: Language Model Improvements

The overall improvements obtained with more advanced language models (size, complexity or amount of data) on a development set are shown in Table 2.

4. ACOUSTIC MODELS

A total of six different models were generated for LVCSR-2000 evaluation.

The primary model, used for all the passes for generating the final rescoring lattices was triphone based (tri), gender dependent model (GD) trained on variance normalized (vn) cepstral parameters. There was one more triphone model used in the final stages which was gender independent (GI), trained on raw cepstral parameters (nvn). This model was originally trained for the LVCSR-98 evaluation. The additional four models used pentaphonic representation (penta), two were gender dependent, the other two gender independent. Two used variance normalized cepstra, the other two used raw cepstra. Each of the model types were built using Vocal Tract Normalized (VTN) [7] [8] cepstral parameters, and Speaker Adaptive Training (SAT) [9]. All of the new models used tied-covariance transformations [**?**] [13]. A total of seven tied covariance transformations was used, two for vowels, and one each for fricatives, nasals, stops, trills+glides, and silence+noises. The labels we use for the six models are: tri_GD_vn, tri_GI_nvn, penta_GD_vn,penta_GD_nvn, penta_GI_vn, penta_GI_nvn.

Tree-based top-down clustering was used for determining state tying. The contextual span of questions used in building the trees define models as triphonic or pentaphonic, depending on whether the questions extended to one phoneme before and one phoneme after or two phonemes before and two phonemes after,

respectively.

Gender dependency implies that one model was trained only on the data from speakers of one gender and the other on the data from speakers of the other gender. Originally, this was true, when we built models on a subset of the training data. We then used those models to determine the gender automatically for each conversation side based on total acoustic likelihood. The same method was also used in the evaluation in the first pass, to determine gender of the speaker in each conversation side (always assuming that there is only one speaker per conversation side). We never measured the accuracy of the automatic gender determination.

Variance normalization of cepstral parameters was performed in the following way. We first defined the target variances for each speaker by finding the average standard deviation across all conversation sides in the training data, using only non-silence frames. These twelve average standard deviations squared (one for each of the twelve cepstral coefficients) became the target variances. For each of the conversation sides we then computed the variances for all non-silence frames, for each cepstral parameter, and scaled each of the parameters so that the new variances match the target variances. The algorithm for determining non-silence frames was very simple: starting from beginning and the end of the utterance (turn) every frame whose energy was below -20.0, up to the first frame whose energy was above -20.0, was considered silence. In addition any frame in the turn whose energy was below -40.0 was also considered silence and was not used in computing the variances. The same algorithm was used in both training and evaluation. This algorithm was compared to using acoustic model provided segmentations to determine which are non-silence frames, and it provided a 0.1% reduction in error rate (absolute) on a development set.

All models also used conversation side based cepstral mean subtraction in training and testing.

In addition to the baseline model used in the first pass to generate word lattices for rescoring (tri_GD_vn) all the models only had two versions: Vocal Tract Normaliation (VTN) trained and Speaker Adaptive Training (SAT) trained models. We first determined VTN factors for each conversation side in the training data using the baseline GD model, and generated cepstral parameters based on those GD VTN factors. In order to do the same for the GI models, we trained a tri_GI_vn model on the minitrain data set (every fifth conversation side from the full training set) and estimated GI VTN factors and generated cepstral parameters based on those GI VTN factors. VTN factors are estimated by performing forced alignment on reference transcription for a conversation side using cepstra generated with the full range of VTN factors. The alignment with the highest likelihood score determined the VTN factor. We used range of 0.9 - 1.1 for GD models and 0.8 - 1.18 for GI models, with increments of 0.02.

The SAT models used VTN models as seed models, and all used Maximum Likelihood Linear Regression (MLLR) adaptation [6], [10] with seven transformations which exactly matched the seven tied-covariance transformations in terms of allocations of gaussians in the acoustic model to one of the transformations. Only the tri_GI_nvn built for the LVCSR-98 evaluation was different, using five transformations in training, with gaussian allocations determined by k-means clustering of gaussian mean values in the model.

All triphone models had approximately 10k states with 12 mixture components each, making a total of approximately 27k three-state left-to-right HMMs. The pentaphone models ranged in size from 12.5k states to 15.5k states with 12 mixture components, making a total of 125k to 155k three-state left-to-right HMMs.

5. LEXICAL MODELS

We primarily used PronLex-based lexicon provided by MSU as available in Nov. 1999. There were about 45k words in the lexicon, but more than 5k were removed from the final wordlist as they were degenerate in some way (spellings trying to match mis-pronunciations, truncated words, etc.) We also added multiple pronunciations to the PronLex system using the methods of the WS97 Pronunciation Modeling team [2], namely:

1. phonetically transcribing the training set w/ an ICSI decision-tree based pronunciation model
2. culling frequently seen alternative pronunciations in these transcriptions and adding them to the baseline lexicon along with their relative frequencies

6. SYSTEM ARCHITECTURE

The recognition system used in the LVCSR-2000 evaluation is based on a cascade of rescoring of word lattices [5] using progressively more accurate models. There are many distinct steps in the recognition process. We will describe the overall strategy first. Later we will define each of the individual steps to complete the description of the system architecture.

The overall system structure can be separated in three steps.

- The first step uses a small acoustic model to generate word lattices of suitable size and accuracy for further rescoring and performs gender determination for each conversation side in the test set.
- The second step uses all of the available models to determine VTN factors, rescores the word lattices and generates output word lattices, one set per model, for further combining into a single set of recognition hypothesis.
- The third step takes the different outputs from the available acoustic models and combines them into a single set of hypothesis as the final recognition result.

Rescoring is the process which adds acoustic and duration costs to the language model costs in the input word lattice and outputs a new word lattice with the combined costs. It is important to remember that each rescoring step in all the experiments consists of two parts: The first part rescores a word lattice which contains 3-gram language model costs by adding acoustic and duration model costs and outputs a word lattice with combined costs; The second part replaces the 3-gram language model costs by 6-gram language model costs and outputs a word lattice with combined costs. If only the best path is required, it is easily extracted from the word lattice that contains the combined acoustic, duration and language costs.

6.1. GENERATION OF WORD LATTICES FOR RESCORING

The first step which generates lattices for rescoring with multiple models consists of several passes.

- The first pass performs recognition with both gender dependent models using a fully composed search network

[3] to output word lattices as finite state machines (FSMs) to be used in rescoring, and to simultaneously perform gender determination. This is an extremely inefficient way to perform gender determination and has been chosen primarily out of convenience. The model with higher total likelihood for the conversation side is selected, defining the gender and word lattices for rescoring. The output word lattices (FSMs) are stripped of all the arc costs, which are then replaced by the 3-gram language model costs. Those are now lattices prepared for use in further rescoring passes. The first pass used model tri_GD_vn to output word lattices. The pruning thresholds were selected so that the lattice word error rate was in the 10-15% range for the *Switchboard* portion of the Dev-98 data set.

- The second pass of the first step finds the best path through the lattices for every speaker/conversation side. They are used with forced alignment in selecting the VTN scale factors. We select the cepstra computed with the chosen VTN scale factor and use them in further rescoring passes with the GD model.
- The third pass rescores the lattices using an acoustic model that was trained on warped data (VTN version of the tri_GD_vn model).
- In the forth pass the recognition result from the third pass (only the best path) is used for alignment with the cepstra and estimation of one global transformation matrix for MLLR adaptation of the SAT version of the tri_GD_vn model. The adapted model is used to once again rescore the word lattices generated in the first pass.
- In the fifth pass we used seven MLLR transformations, with Gaussian mean classes exactly matching the tied-covariance transformation classes to adapt the model. The recognition run in this pass (not a rescoring pass) using the adapted SAT tri_GD_vn model is used to generate even sparser and more accurate word lattices for further rescoring steps.

Those word lattices are the final output of the first step and will be used for rescoring with all of the six acoustic models in the second and third steps. The additional step of generating rescoring lattices with the adapted model reduces the error rate by 0.3% with the LVCSR-98 system on Eval98 by reducing search errors. It also provides significantly smaller lattices which improves recognition speed and memory usage.

6.2. GENERATION OF WORD LATTICES BY MULTIPLE MODELS

The second step essentially mimics the second through fifth passes of the first step, using the VTN and SAT versions of the six available models. There is only one exception. The model tri_GI_nvn continues to use one MLLR transformation in the fifth pass as it does not use tied-covariances. The result of the second step is a set of word lattices with combined acoustic, duration and 6-gram LM costs for each of the six available models.

6.3. MODEL COMBINATIONS FOR IMPROVED RECOGNITION ACCURACY

The third step of the recognition process uses the lattices generated with the six different models to generate a single combined hypothesis.

This is achieved in several passes.

	Word Error Rate (%)
best single model	36.90
combined output	35.41
best single model (combined adaptation)	35.75
combined output (combined adaptation)	34.93

Table 3: Combined model improvements

- The first pass combines the lattices by intersection. First we order the models by expected word accuracy, from most accurate to least accurate. Next we intersect all the lattices generated by the first two models. If any of the intersections result in an empty lattice (there were no common paths in the outputs of the two models) than the first model lattice is kept (or the result of the previous intersection). The result of this process is then used to intersect with the output lattice of the next model, and so on, until all six model outputs have been used. This process is essentially the same to the process used in the LVCSR-98 evaluation.
- The second pass is new in our system. It uses the one-best outputs of all the models and the combined output lattice (total of seven recognition hypothesis) as the reference output for MLLR adaptation, with all of the hypothesis having the same weight. The intuition is that the correctly recognized words will be repeated seven times, while the mis-recognitions will be randomly different, and thus will get lower weight, in essence providing a version of a confidence score. We then use the adapted models (seven MLLR transformations, one MLLR transformation for tri_GI_nvn) to yet again rescore the lattices. We also perform LM rescoring with the adapted version of the 6-gram LM in this pass.
- The third pass performs the same function as the first pass on the resulting lattices from the second pass. The best path from the combined output in this pass provides the final recognition output.

On a subset of the eval-98 data we obtained the results shown in Table 3.

7. LVCSR-2000 RECOGNITION RESULTS

The recognition results for different passes in the first step of the recognition process are shown in Table 4. It shows word error rate separately for *Switchboard* and *Call Home* parts of the evaluation data, as well as the combined word error rate. The

Word Error Rate (%)				
Model/pass	LM	CH	SWBD	Combined
baseline	3-gram	43.8	31.0	37.4
baseline	6-gram	41.9	29.8	35.9
VTN	3-gram	42.1	29.3	35.8
VTN	6-gram	40.4	27.6	34.0
MLLR-1	3-gram	40.1	27.1	33.6
MLLR-1	6-gram	38.6	25.8	32.2
MLLR-7	3-gram	38.2	25.5	31.9
MLLR-7	6-gram	37.9	25.3	31.6

Table 4: The first step of the recognition process

Word Error Rate (%)							
Model/pass	LM	penta_GD_vn	penta_GD_nvn	penta_GI_vn	penta_GI_nvn	tri_GD_vn	tri_GI_nvn
VTN	3-gram	33.7	33.4	34.3	34.4	34.6	35.7
VTN	6-gram	32.2	32.1	32.9	33.0	32.2	34.2
MLLR-1	3-gram	31.7	31.7	32.4	32.3	33.1	33.6
MLLR-1	6-gram	31.1	30.5	31.1	31.0	31.8	33.0
MLLR-7	3-gram	30.5	30.4	31.1	31.0	31.7	33.6
MLLR-7	6-gram	30.3	30.2	30.8	30.7	31.4	32.6

Table 5: The second step of the recognition process

Word Error Rate (%)							
Model/pass	LM	penta_GD_vn	penta_GD_nvn	penta_GI_vn	penta_GI_nvn	tri_GD_vn	tri_GI_nvn
MLLR-7	6-gram	30.3	30.2	30.8	30.7	31.4	32.6
combined	6-gram		29.6	28.9	28.8	28.7	28.6
MLLR-7-combined	6-gram	29.5	29.6	29.9	29.9	30.1	33.4
combined	6-gram		28.9	28.5	28.5	28.5	28.4
combined	adapt. 6-gram		28.8	28.4	28.5	28.4	28.4

Table 6: The third step of the recognition process

results are shown for the first pass with the baseline model, VTN model result, and results for the adapted model results, both when using one and seven MLLR transformations. For each pass there is also a result when a 3-gram and 6-gram language models are used.

The recognition results for different passes in the second step of the recognition process are shown in Table 5. It shows the combined word error rate for each of the six different models in all of the different passes of the second step for both the 3-gram and 6-gram language models.

The recognition results for different passes in the third step of the recognition process are shown in Table 6. It shows combined word error rate for each of the six different models in all of the different passes of the third step including the result of combining all the model outputs. It separates the single model accuracy when the model is adapted only on the previous best output of that model and when it is adapted on all the best outputs of all the models plus the best combined output. Only the results with the 6-gram language model are shown. First the results with the standard 6-gram LM are shown, and at the end the results with the adapted version of the 6-gram LM are shown. Also, for the rows marked "combined" the result is based on combining all the models in the columns to the left of that column and the model in that column from the previous row in the table. The first column is empty to indicate there was no combining of model outputs, the second column is the combination of the first and second model outputs, the third column is the combination of the first, second and third model outputs etc. The results generally get better as more model outputs are combined. The row marked MLLR-7-combined shows results for the single model which was adapted using outputs of all the models and their combined output. The first row is just a copy of the last row from Table 5 for comparison.

It is important not to observe any of the rescoring results in isolation. There is an interaction between the model used to generate rescoring lattices, transcription used to perform adaptation of the acoustic model and the model used to perform rescoring, that can significantly affect recognition results. For example, the VTN result for tri_GD_vn in step one and step two (35.8 and 34.6 with 3-gram LM; 34.0 and 32.9 with 6-gram LM respectively) are significantly different, and yet the only difference was in the lattices they were rescoring, even though the lattices were generated with different versions of the same tri_GD_vn model. The bigger the change in the setup the stronger the effects on the error rate. It can also be seen in the differences between the GD and GI results for otherwise comparable results. In the experiments on the development set we generally found GD models to have the error rate about 2.0% absolute lower than GI models. The difference is much lower in the rescored results as the GI score is artificially improved by rescoring relatively sparse rescoring lattices generated by a GD model.

8. CONCLUSION

We have presented our LVCSR-2000 recognition system for use in the DARPA evaluations on the *Switchboard* and *Call Home* speech corpora. The system consists of an initial recognition step which outputs word lattices which are then rescored using a large number of acoustic models which were deliberately trained to maximize their differences. The final step uses two different ways to combines the outputs by the different acoustic models to improve recognition performance. The more conventional way intersects different model generated word lattices for improving recognition accuracy. The new method uses the best paths from different models as reference transcriptions for performing the MLLR adaptation and improving the adaptation process, similar to what can be achieved by using confidence scores. When those outputs are combined by intersecting output word lattices we achieve significantly improved recognition results when compared to any of the individual models alone.

9. REFERENCES

1. K. Seymore and R. Rosenfeld. Scalable backoff language models. In *Proceedings of ICSLP*, Philadelphia, Pennsylvania, 1996.

2. W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos. Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In *Proc. ICASSP '98*, Seattle, WA, 1998.
3. M. Mohri, M. Riley, D. Hindle, A. Ljolje, and F. Pereira. Full expansion of contex-dependent networks in large vocabulary speech recognition. In *Proc. ICASSP '98*, Seattle, WA, 1998.
4. A. Ljolje, M. Riley, and D. Hindle. The AT&T Large Vocabulary Conversation Speech Recognition System. In *Proc. Eurospeech '99*, Budapest, Hungary, 1999.
5. A. Ljolje, F. Pereira, and M. Riley. Efficient General lattice Generation and Rescoring. In *Proc. Eurospeech '99*, Budapest, Hungary, 1999.
6. V. Digilakis, D. Rtischev, and L. Neumeyer. Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, no. 5, pp. 357-366, 1995.
7. L. Lee, and R. Rose. Speaker Normalization using Efficient Frequency Warping Procedures. In *Proc. ICASSP '96*, Atlanta, GA, 1996. pp. 353-356, 1996.
8. T. Kamm, G. Andreou, and J. Cohen. Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. In *Proc. 15th Annual Speech Research Symposium*, pp. 161-167, CLSP, Johns Hopkins University, Baltimore, MD, June 1995.
9. T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul. A Compact Model for Speaker-Adaptive Training. In *Proc ICSLP '96*, Vol.2, pp. 1137-1140, 1996.
10. C. Leggetter, and P. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. *Computer Speech and Language*, Vol. 9, No. 2, pp. 171-186, 1995.
11. A. Ljolje. The Improtance of Cepstral Parameter Correlations in Speech Recognition. *Computer Speech and Language*, Vol. 8, pp. 223-232, 1994.
12. M. Gales. Semi-tied covariance matrices. In *Proc. ICASSP '98*, pp. 657-660, 1998.