PERFORMANCE OF THE IBM LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM ON THE ARPA WALL STREET JOURNAL TASK

L.R. Bahl, S. Balakrishnan-Aiyer, J.R. Bellgarda, M. Franz, P.S. Gopalakrishnan
D. Nahamoo, M. Novak, M. Padmanabhan, M.A. Picheny, S. Roukos
IBM T. J. Watson Research Center
P. O. Box 704, Yorktown Heights, NY 10598

ABSTRACT

In this paper we discuss various experimental results using our continuous speech recognition system on the Wall Street Jounal task. Experiments with different feature extraction methods, varying amounts and type of training data, and different vocabulary sizes are reported.

1 INTRODUCTION

Large vocabulary continuous speech recognition is an area that is of great current interest, and to this end, several speech recognition systems have evolved that are capable of dealing with such recognition tasks [2, 4, 5, 6, 7, 9]. The ARPA sponsored Wall Street Journal task represents a standardized database that enables the evaluation of the features specific to these different systems on a common platform. In this paper, we present the performance of the IBM continuous speech recognition system on this task. We will concentrate on the speaker-independent portion of the database. The test data used in the experiments is read speech recorded using a Sennheiser microphone. We report experimental results showing the influence of several parameters of the system, including different ways of extracting feature vectors, varying acoustic and language modelling vocabulary sizes, and different amounts and types of training data on the recognition performance.

The rest of this paper is organized as follows: in the next section, a brief description of the IBM system is given. In Section 3 a number of experiments are described, along with their results, that show the interaction between system parameters and the recognition performance, and finally, conclusions are given in the last section.

2 SYSTEM DESCRIPTION

Essential aspects of the system used in the experiments here have been described earlier [1, 2]. We summarize below the important elements of the system and point out the enhancements that have been introduced.

Signal Processing: Acoustic feature vectors are extracted from the 16KHz sampled data every 10 ms. The processing involves (i) computing 24-band mel cepstra every 10 ms using a 25 ms window for the FFT, (ii) splicing together the cepstra from the adjacent s frames on either side of the current frame (typically $s = 4$) (iii) applying a transformation that brings the dimensionality of the vector down to, say, 60 dimensions (iv) outputting the 60-dimensional vector as the feature vector at the current frame.

Linear discriminant analysis [2] is used to compute the above transformation from training data. In this system, the procedure is modified in the following manner. This technique essentially is a two-step process for extracting linear discriminants; in the first step, the linear discriminants of the unspliced 24 dimensional cepstra are obtained, and applied on the cepstra - there is no reduction in dimensionality at this stage. The second step of the double rotation technique attempts to capture the dynamics of speech in this transformed 24-dimensional space. This is done independently for each dimension, d, of the transformed space by splicing together the d^{th} component of the transformed cepstra, across $2s+1$ frames, and obtaining linear discriminants to maximally separate subphonetic classes on the basis of this $(2s+1)$-dimensional vector. Subsequently, the 60 most discriminative projections are chosen and put together with the first rotation to give the final transformation.

The feature extraction technique described here is in contrast to that used in several other systems [5, 9], which use differences in the cepstral vectors between frames to model the dynamics of speech. The results of the experiments in the following section indicate that our method provides a performance improvement.

0-7803-2431-5/95 $4.00 © 1995 IEEE

Search: The IBM system uses an envelope search algorithm [3]. An important feature of this technique is that it does not rely on the language model to limit the size of the search space, unlike Viterbi techniques [5, 9, 6, 4, 7]. Consequently, it is possible to increase the size of the acoustic vocabulary beyond the size of the language model vocabulary. This will be elucidated further in the next section.

Acoustic Models: The system uses different acoustic models for sub-phonetic units in different contexts. These instances of context dependent classes are identified by growing a decision tree from the available training data [1] and specifying the terminal nodes of the tree as the relevant instances of these classes. The acoustic feature vectors that characterize the training data at the leaves are modelled by a mixture of Gaussian pdf's, with diagonal covariance matrices. The HMM's used to model the leaves are simple 1-state models, with a self-loop and a forward transition. As far as the output distributions on the state transitions of the model are concerned, rather than expressing the output distribution directly in terms of the feature vector, by using the mixture of Gaussian pdf's modelling the training data at the leaf [5, 9], the IBM system expresses it in terms of the rank of the leaf [2]. The rank of a leaf is obtained by computing the log likelihood of the acoustic vector using the model at each leaf, and then ranking the leaves on the basis of their scores.

Training of the acoustic models: The training procedure assumes that we have an initial speaker independent training that can be used to bootstrap the procedure. We begin by marking, for each training utterance, the particular pronunciation of each word and also the presence of silence between words. This is done by decoding the training utterances using the initial training, but constraining the decoder to consider only different pronunciations of each word and silences between words. The training data is then aligned against these scripts using the Viterbi algorithm, giving us the class label corresponding to each feature vector (transition within the phonetic HMM to which the vector is aligned) and the phonetic context. This information is first used to compute the matrices for 2-level LDA described earlier. Decision networks are then constructed using the training data, one for each class label. The training data at each terminal node is then used to determine a mixture of Gussian densities with diagonal covariances. This is done by first clustering the feature vectors and then refining the models using the forward-backward training algorithm. Optionally, this whole procedure may be repeated after separating the data corresponding to each gender (including the estimation of the projection matrices) to yield gender dependent models. The number of mixture components is variable and an upper bound is often imposed by the amount of training data available. We will see the effect of these factors on recognition accuracy in the next section.

Sentences	Components	% Error	% Improvement
6850	31653	9.69 %	
14257	54823	8.33 %	14 %
21385	74666	8.17 %	1.9 %
28513	90434	8.12 %	0.62 %

Table 1. Effect of varying amount of training data

Language Model: The standard language model used in all the experiments is a trigram model. Different amounts of training data is used for deriving the models for different vocabulary sizes. This is covered in greater detail in the next section.

3 EXPERIMENTAL RESULTS

We conducted several experiments on the speaker independent portion of the ARPA Wall Street Journal corpus, as well as the NAB news test data that was available in 1994. The training data consists of read speech recorded using a high quality Sennheiser microphone. We distinguish between two training sets - a collection of roughly 35,000 sentences from 284 speakers labeled as *SI-284* and another collection of sentences of similar size from 37 different speakers labeled *SI-37*. Note that *SI-284* contains a larger variety of speakers but smaller amount of data from each speaker, whereas *SI-37* contains a smaller variety of speakers but significantly larger amount of data from each speaker. Three different sets of test data are used in the experiments below. The data labeled ***dev-92*** consists of 403 sentences of test data from ten different speakers, that was used as the development test data in 1992. The data labeled ***eval-93*** consists of 213 sentences of test data from ten different speakers, that was used as the benchmark test of systems in November 1993 [8]. The test data labeled ***dev-94*** consists of 310 sentences from 20 speakers that was available in September 1994. This data comes from the North American Business News (NAB) domain.

We first examine the effect of varying the amount of training data. Experiments were conducted using the ***dev-92*** test set and various amount of training data from *SI-284*. Table 1 shows the error rate as a function of the amount of training data. The first column shows the number of sentences of training data. Note that the number of mixture components changes with the amount of training data, since upper limits are placed on the number of components in order to obtain reliable estimates of the parameters of the Gaussian distributions. The third column shows the error rate and the last column shows the percentage improvement in error rate in each step. The feature vectors used here are 12 mel cepstra augmented with the overall energy, and first and second order differences of these 13 parameters.

We next contrast the feature vectors derived using the two-level LDA scheme described in the previous section with the feature vectors used in the first experiment, namely, cepstra, energy, and first and second order differences. Table 2 shows the error rate using these two techniques. The *eval-93* test set was used for this experiment. Training in both cases used all sentences from *SI-284*, with the systems being gender independent, and with 124K mixture components. As can be seen, using the feature vectors obtained through the 2-level LDA provides a small increase in the accuracy. This difference is further accentuated for the gender dependent case, where the rotation matrices can be specifically made for each gender.

As mentioned earlier, the maximum number of mixture components that can be used is limited by the requirement that reasonable estimates of these parameters can be obtained from the limited amount of available training data. The effect of varying the number of mixture components while keeping the amount training data constant is indicated in Table 3. The feature vectors in this case had 39 dimensions, and were derived using the 2-level LDA technique. The *eval-93* test data was used for the experiment. All sentences from *SI-284* were used for training the systems.

All experiments reported above were performed using a gender independent training set. It is possible to segregate the data depending on the gender and train two different sets of models. Gender selection can then be made by decoding three different ways (two genders and the gender independent) and selecting the sentence with the highest likelihood. Table 4 compares the word error rate obtained using such a gender dependent system to a gender independent system on the *eval-93* test set.

A further advantage associated with using the 2-level LDA technique to extract feature vectors is that the dimensionality of the vector can be changed easily. As one might expect, increasing the number of dimensions does help, but only upto a point (performance starts degrading after 60 dimensions). The results of these experiments on the *eval-93* data are given in Table 5. The systems were gender dependent, and trained on all the *SI-284* data.

In the next experiment (Table. 6), we present results on the *dev-94* data, obtained by training a gender dependent system on *SI-37* data[1], and compare it to results obtained by training it on *SI-284* data (17662 sentences from male speakers, and 17996 sentences from female speakers). The results show that having a larger number of speakers in the training corpus gives better speaker-independent performance.

As mentioned earlier, in the envelope search technique, the vocabulary used for the acoustic models is independent of the language model vocabulary. This is because the candidates for further extension of a path are obtained by searching the acoustic vocabulary using simple models. Hence, even if the size of the language model vocabulary is constrained, it is possible to have a larger acoustic vocabulary [2]. This in turn reduces the problem of out-of-vocabulary words, as these words may now appear as candidates for extension after the fast match. As far as the language model is concerned however, all the words proposed for extension by the fast match that are not in the language model vocabulary are treated as the "unknown" word, and assigned the "unknown" word probability. We present in Table 7, the results of decoding the *dev-94* data using a gender independent system with acoustic vocabularies of 20K and 64K respectively. A 20k trigram language model is used in both cases. Also, the system has 124K mixture components, and uses 50-dimensional acoustic feature vectors, that are obtained using the double rotation technique.

Finally, we present experiments that demonstrate the effect of the size of the lexicon. The lexicon here refers to both the acoustic vocabulary and the language model vocabulary. Results are presented in Table. 8 for the *dev-94* data, using a gender dependent system with 120k mixture components, that uses a 60 dimensional feature vector obtained using the double rotation technique. The lexicon sizes were respectively 20K and 64K, and the percentage of words in the test data outside the vocabulary decreases from 2.7% for the smaller vocabulary to 0.7% for the larger vocabulary. The large gains possible by increasing the lexicon size are further illustrated by comparing the performance on the *eval-93* data; the 64K system gives a word error rate of 7.9% on the data compared to 10.3% using the 20K vocabulary.

[1] Actually, data from only 25 out of the 37 speakers was used. This data comprised of 13372 sentences from male speakers and 15068 sentences from female speakers.

[2] In contrast, viterbi-based techniques use the language model to predict a shortlist of candidate words for extension, hence there is no advantage that accrues from having an acoustic vocabulary that is larger than the language model vocabulary.

4 CONCLUSIONS

In this paper we described some experiments with the IBM large vocabulary continuous speech recognition system on the ARPA task.[3] It is clear from the results that to obtain good speaker independent performance, we need to use a large amount of training data from a large number of speakers. It possible to obtain more accurate models in a higher dimensional feature space. Though gender dependent systems appear to provide better accuracy, we believe that this is not a very realistic arrangement for practical systems (besides being politically incorrect).

Cepstra	2-level LDA	% Improvement
11.86 %	11.52 %	2.9 %

Table 2. Cepstral parameters versus two-level LDA

90K compnts	124K compnts	% Improvement
12.5 %	11.52 %	7.8 %

Table 3. Effect of number of mixture components

Gender indep.	Gender dep.	% Improvement
11.52 %	10.9 %	5.4 %

Table 4. Gender independent versus dependent training

40-dim LDA	60-dim LDA	% Improvement
10.9 %	10.3 %	5.5 %

Table 5. Effect of number of dimensions

$SI - 37$	$SI - 284$	% Improvement
14.26 %	12.08 %	15.3 %

Table 6. Effect of training on $SI - 37$ vs. $SI - 284$

20K vocab	60K vocab	% Improvement
12.2 %	11.1 %	9 %

Table 7. Effect of baseform set size

20K vocab	60K vocab	% Improvement
11.6 %	9.4 %	19 %

Table 8. Effect of lexicon size

REFERENCES

[1] L. Bahl, P. de Souza, P. Gopalakrishnan, M. Picheny, "Context-dependent vector quantization for continuous speech recognition," Proc. IEEE ICASSP-93, Minneapolis, MN, May 1993.

[2] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny, "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. Intl. Conf. Acoust., Speech and Sig. Proc., 1994.

[3] L. R. Bahl, P. S. Gopalakrishnan, R. Mercer, "A tree search strategy for large vocabulary continuous speech recognition", elsewhere in these proceedings.

[4] V. Digalakis, H. Murveit, "An algorithm for optimizing the degree of tying in a large vocabulary hidden markov model based speech recognizer", Intl. Conf. Acoust., Speech and Sig. Proc., April, 1994.

[5] J. L. Gauvain, L. F. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task", Intl. Conf. Acoust., Speech and Sig. Proc., April, 1994.

[6] M. Hwang, R. Rosenfeld, E. Thayer, R. Mosur, L. Chase, R. Weide, X. Huang, F. Alleva, "Improving speech recognition performance via phone dependent VQ codebooks and adaptive language models in Sphinx-II", Proc. Intl. Conf. Acoust., Speech and Sig. Proc., April, 1994.

[7] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz, G. Zavaliagkos, "Comparative Experiments on Large Vocabulary Speech Recognition", Intl. Conf. Acoust., Speech and Sig. Proc., 1994, pp I-561-564.

[8] D. S. Pallett et. al., "1993 benchmark tests for the ARPA Spkoken Language Programs", Proceedings of ARPA Speech and Natural Language Workshop, pp 51-73, 1994.

[9] P. C. Woodland, J. J. Odell, V. Valtchev, S. J. Young, "Large vocabulary continuous speech recognition using HTK", Intl. Conf. Acoust., Speech and Sig. Proc., April, 1994.

[3] All experiments were run on non-Pentium machines