HMM Training Based on Quality Measurement

Yuqing Gao, Ea-Ee Jan, Mukund Padmanabhan, Michael Picheny
IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY
email: yuqing,ejan,mukund,picheny@watson.ibm.com

Abstract

Two discriminant measures for HMM states to improve effectiveness on HMM training are presented in this paper. In HMM based speech recognition, the context-dependent states are usually modeled by Gaussian mixture distributions. In general, the number of Gaussian mixtures for each state is fixed or proportional to the amount of training data. From our study, some of the states are "non-aggressive" compared to others, and a higher acoustic resolution is required for them. Two methods are presented in this paper to determine those non-aggressive states. The first approach uses the recognition accuracy of the states and the second method is based on a rank distribution of states. Baseline systems, trained by a fixed number of Gaussian mixtures for each state, having 33K and 120K Gaussians, yield 14.57% and 13.04% word error rates, respectively. Using our approach, a 38K Gaussian system was constructed that reduces the error rate to 13.95%. The average ranks of non-aggressive states in rank lists of testing data were also seen to dramatic improve compared to the baseline systems.

I. Introduction

In speech-recognition applications, the size of the acoustic model is closely related to both the system performance and the computational resource requirement. In general, if sufficient training data is available, one would try to use as many model parameters as possible to achieve better recognition accuracy. The experimental results also show that the more model parameters used, the better the performance. However as the number of parameters increases, more computational resources, (typically larger memory footprint and more computation,) are required.

In conventional Large Vocabulary Continuous Speech Recognition (LVCSR) systems, words are represented as sequences of phones. Each phone is modeled by a several states HMM. A decision tree to determine the context information is then constructed from the training data and the terminal nodes of the tree, representing collections of instances of these classes, are grouped according to context. These classes, also called context-dependent states or leaves, are modeled by a mixture of Gaussian pdf's with diagonal covariance matrices. Thus, the model parameters are mainly means, variances and prior distributions of the mixture of Gaussian components. In general, if sufficient training data is available, modeling these classes with more Gaussians yields better performance. Conventionally, the number of Gaussian mixtures for a class is fixed or is chosen to be proportional to the number of data samples, with some loose constraints, for example, the amount of training data must be sufficient to robustly estimate the parameters of the components. This kind of approach may result in models not optimized as far as the recognition error rate is concerned.

We analyzed the overall system errors and took a close look at the contribution of each of these states to the recognition errors, and noticed that not all of them are equally important, i.e., not all states contribute equally to the errors. Some states are more often misclassified as others, and some states often encroach upon the space of other states. We refer to these two conditions as "non-aggressive" and "invasive" states [1].

Our strategy is to choose the number of mixture components to increase the number of mixtures for those "non-aggressive" states that contribute more to the misclassification, and also to decrease the number of mixtures for "invasive" states which are encroaching on others' spaces.

We associate a quality measurement with each state, that reflects the degree of non-aggressiveness or invasiveness of the state, and that therefore enables us to optimize the number of mixtures and the performance of HMMs for each state separately. The overall system size can hence be optimized given a fixed amount of the training data.

This model complexity issue has been previously addressed[1, 2]. In this paper, two different approaches have been explored to measure the quality of context-dependent classes to improve effectiveness of the acoustic modeling. The experimental results using these measures to control the size of the acoustic models are also presented and show that both methods are very effective.

II. Recognition Accuracy Based State Quality Measurement

The first method we describe in this paper is to use the classification accuracy of a state as the quality measure. This classification accuracy of a state is obtained by:

1. Decoding a large number of training speech sentences and Viterbi aligning the the training speech against the decoded transcription;

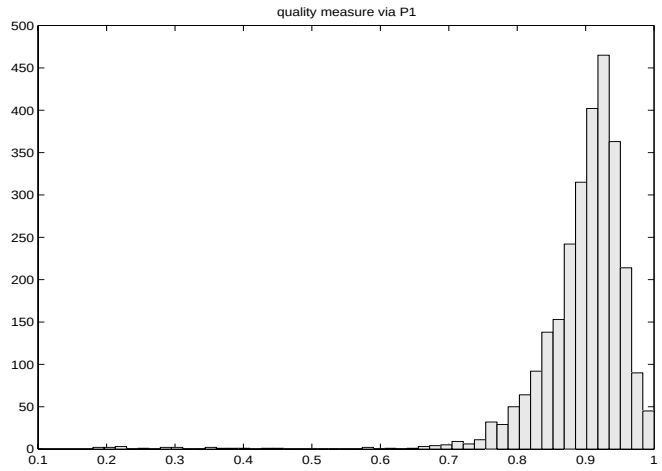

Figure 1: Accuracy based quality measure distribution

2. Viterbi aligning the training acoustic data against the correct transcription;

3. Tagging each analyzed speech frame with 2 state ids, one associated with the correct path, and the other with the decoded path;

4. Let C_i = total number of frames tagged as state i in the correct path; D_i = number of frames correctly tagged as i in the decoded path.

Obviously, $D_i <= C_i$. Therefore, the accuracy based confidence measure for state i is defined as:

$$P_{1i} = \frac{D_i}{C_i} \tag{1}$$

P_{1i} is the correct probability of state i. P_{1i} measurement can be considered as a special case of P_c^l of [1], which is obtained by decoding the training data and producing N-best lists of hypotheses, subsequently, Viterbi aligning against each hypothesis. In this special case, only the best decoding path is used for estimating P_{1i}. In the ideal case, the decoding path would be the same as the correct path, therefore P_{1i} should equal to 1 for all state i. However, the ideal case is rarely achieved, and in general states have values of P_{1i} lying between 0 and 1. In figure 1, we showed a histogram of P_{1i} which is obtained from one of our baseline system, which has 2755 states, and each modeled by at most 12 Gaussians, the total number of Gaussians being 32626.

We set a threshold of 0.80 for P_1, i.e., if $P_{1i} < 0.80$, it implies that all states which have P_{1i} less than 0.80 are considered "non-aggressive" states and more Gaussian mixtures should be used to model them. All states which have P_{1i} equal or greater than 0.80 are considered "good" or in some sense "aggressive" states and the number of Gaussian mixtures for them can be left unchanged or can be reduced.

III. Rank Based State Quality Measurement

IBM's large vocabulary, continuous speech recognition system is a rank based system[3]. In a rank based system, for each feature vector all state likelihoods are computed and they are ranked in the order of likelihoods. If the correct state corresponding to a feature vector has a low rank, it will probably lead to a decoding error. The rank based state quality measurement is defined according to the ranking of a state.

Besides, the measurement is purely based on acoustics, as we believe that if language models are applied during decoding as in P_{1i} and in [1], the state score benefits from the LM score, therefore it prevents us from finding the true goodness of the acoustic model of a state. Even though we can eliminate the effect of the language models by disabling the LM during decoding, the decoding path still limited by the context and the search space. The rank based confidence measure tries to address the problem purely from using acoustic scores.

In order to obtain the rank based state quality measurement, we follow the procedure:

1. For each frame in training data, find the top N ranked states.

2. Let R_i equal the number of frames that contain the correct state i in the top N state list.

3. Let C_i equal the total number of frames tagged as state i in the correct path.

Clearly R_i is always equal or less than C_i. The definition of rank based state quality measurement is:

$$P_{2i} = \frac{R_i}{C_i} \tag{2}$$

Some examples of this quality measure are showed in Figure 2. The system used to compute the measurement histogram is the same as for Figure 1. As it can be seen in Figure 2, different N's for top ranks result in different distributions of the quality measurement. For example, when N=30, it means that only if, for a speech sample x(t), the correct state is in the rank list of top 30, the sample x(t) is considered correct during the quality measurement computing. The smaller N of the top rank list, the stricter condition is required for being a correct state. When we set a threshold 0.80 for P_{2i}, i.e., all states which have P_{2i} less than 0.80 are considered "non-aggressive" states and more Gaussians should be used because the probability of the correct state in the top N-best rank list is less than 0.80. Those states which have P_{2i} equal or greater than 0.80 are considered "good" or in some sense "invasive" states and the number of Gaussian mixtures for them can remain unchanged or can be reduced.

The size of Top N rank list together with the threshold for dividing "non-aggressive" and "invasive" states is determined from experiments.

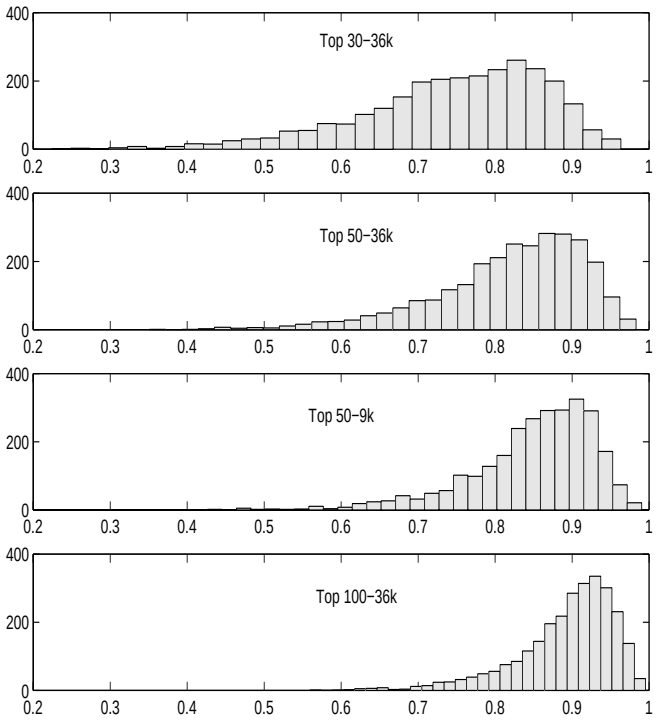

Figure 2: Rank based quality measure distribution

IV. Experimental Results

The training data is 80-hours of in-house speech data (36,000 sentences). The decision tree, with 2755 states (leaves), has being used in all experiments. 3 baseline systems, where each state has a maximum of 12, 19 or 60 Gaussians respectively, are trained with this training data. The total number of Gaussians in these 3 systems has 33K, 51K and 120K respectively.

The test script is an in-house office correspondence script, which includes 61 sentences, with 1117 words, read by four males and six females in a quiet office using an ANC500 headset microphone.

The signal processing uses a 16KHz sampling rate and extracts 13 dimension mel-scale cepstra (with C0) with their first and second order derivatives every 10ms.

The recognition word error rates for the 3 baseline systems are 14.57%, 13.91% and 13.04%, respectively. Our goal is to make new systems using the P_1 or P_2 measurement to control the number of Gaussian mixtures for each state, and achieve a lower word error rate for a given number of Gaussians.

A. Experiments with P_1

The P_1 distribution in Figure 1 is produced by decoding all 36K training sentences using the 32K Gaussian baseline system.

Two new systems were built with the P_1 state quality measurement. In the $N_1 8 - N_2 26$ system, a threshold of 0.88 is used for P_1. That makes 774 out of 2755 states become "non-aggressive" states because their P_1's are less than 0.88. We use a maximum of 26 Gaussians for the "non-aggressive" states, while a maximum of 8 Gaussians for those states whose P_{1i} are equal to or greater than 0.88. The resulting system has a total of 32K. The word error rate for this new system is 14.22% which is lower than the Baseline 32K Gaussian system.

# of Gaussians	Baseline	P_1
32K	14.57% N=12	14.22% N_1=8, N_2=26 th=0.88, (774)
51K	13.91% N=19	13.46% N_1=12, N_2=28 th=0.91, (1378)
120K	13.04% N=60	-

Table 1: Baseline and P_1 systems

The second system $N_1 12 - N_2 28$ was targeted to have 51K Gaussians. A threshold of 0.91 is used, that produces 1378 "non-aggressive" states. We used a maximum of 28 Gaussians for these "non-aggressive" states and 12 Gaussians for the rest of the states. The recognition error rate of the system is also lower than the 51K baseline system. The comparison results using this approach is illustrated in Table1.

B. Experiments with P_2

We experimented with different sizes of N-best rank list, thresholds for dividing "non-aggressive" and "invasive" states, as well as the amount of training data needed to generate P_2 distribution for evaluating the rank based state quality measurement.

In Table2, the size of the N-best rank list was fixed at 50, while the amount of training speech used for generating P_2 distributions are 36K and 9K sentences, respectively. The P_2 distributions are in Figure 2. Both systems in Table2 have the same number of total Gaussians (38K). Although the thresholds used are different (0.82 vs. 0.85), the number of "non-aggressive" states are about same (1180 vs. 1141 states), therefore the maximum numbers of Gaussian mixtures for "non-aggressive" and for "aggressive" states are same for 2 systems: 8 and 23, respectively.

According to the recognition word error rates showed in Table2, the system obtained from analyzing 36K training data is slightly better than the system obtained from analyzing 9K training data, but the difference is small. It can be concluded that the P_2 distribution is not really sensitive to the amount of training data used to estimate it.

In Table3, we compare the effect of the size of the N-best rank list and different number of Gaussian mixtures for each state. Three different non-aggressive state lists were calculated using top N of 30, 50 and 100, respectively.

# of Training data	top50-36k	top50-9k
# of mixtures per state	N_1=8, N_2=23	N_1=8, N_2=23
Threshold	0.82 (1180)	0.85 (1141)
Error rate	13.95%	14.06%

Table 2: Error rate on quality measure via size of training data. Both systems are composed by approximately 38K Gaussians

top N best of Rank distribution	Systems with 38K Gaussians	
30	14.07% N_1=8, N_2=23 th=0.74 (1167)	14.33% N_1=8, N_2=55 th=0.63, (459)
50	13.95% N_1=8, N_2=23 th=0.82 (1180)	14.17% N_1=8, N_2=55 th=0.73, (455)
100	14.32% N_1=8, N_2=23 th=0.89 (1049)	-
baseline (33K)	14.57%	
baseline (120K)	13.04%	

Table 3: Performance Comparison using difference sizes of N-best rank list and different numbers of Gaussian mixtures per leaf by rank based state quality measurement. The systems are targeted at approximately 38K Gaussians.

Three systems are then constructed using these lists, 23 Gaussians for non-aggressive states and 8 for the others. (N_1=8, N_2=23 in Table3). It is found that the system via the top 100 list is worse than two others, and the system using the top 50 is slightly better than the one from top 30 although the difference may not be significant. We checked the ranks of the correct states in the rank list, the average rank is between 10-20. It implies that using the top 100 list is not a favorite choice.

From the second and third rows in Table 3, it shows that for a target system size (38K), a system with more non-aggressive states with slightly larger Gaussian mixtures (approximately 1100 states with 23 Gaussians each) is better than one with less non-aggressive states with much larger number of mixtures (approximately 500 with 55 Gaussians each).

Systems are evaluated from 2 different perspectives. One obvious way is to use the system recognition error rate, the another way is to see the improvement in rank of the correct state on test data. In Table4, for each condition of different sizes of N-best rank lists, we compare the average rank of "non-aggressive" states in the rank lists over 3 different systems: the N12 (32K Gaussians) baseline system, the $N_18 - N_223$ system, and the N60 (120K Gaussians) baseline system. The average ranks in N60 system are higher than in the N12 system as expected. The average ranks in our $N_18 - N_223$ system are also higher

system	Average Ranks		
	N=12	N_1=8, N_2=23	N=60
top 30	21.96	19.64	18.70
top 50	22.05	19.77	18.82
top 50 (9k)	21.77	19.48	18.59

Table 4: Average Ranks improvement over different systems

than in N12 system and closer to N60 system as desired. This reflects the advantages of our method.

V. Conclusion

We proposed two approaches to determine the non-aggressive states in HMM based speech recognition systems. Instead of assigning a fixed number of Gaussian mixtures for all states, larger number of Gaussians are used for those non-aggressive states and smaller number of Gaussians for the rest of the states. The accuracy can be improved without increasing the system size. Both approaches yield comparable improvement. From our experimental results we also conclude that for a target system size, a system with more non-aggressive states with slightly larger Gaussian mixtures is better than one with less number of non-aggressive states with much larger number of mixtures.

References

[1] L. R. Bahl, M. Padmanabhan, "A Discriminant Measure for Model Complexity Adaptation", Proceedings of the ICASSP, pp 453-456, 1998.

[2] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training", Proceedings of the ICASSP, pp 449-452, 1995.

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