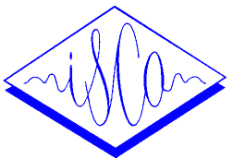

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

VOCABULARY-BASED ACOUSTIC MODEL TRIM DOWN AND TASK ADAPTATION

Guo Qing[†], *Yan Yonghong, Yuan Baosheng, Zhang Xiangdong, Jia Ying and Liu Xiaoxing*

Email: baosheng.yuan@intel.com
Intel China Research Center, Beijing, PRC

ABSTRACT

In this paper, a vocabulary trim down algorithm is proposed in decision tree-based acoustic model to make the model more close to the given task. Using this trim down model as seed model to do task adaptation is also presented. Based on this framework, users can configure the acoustic model by themselves according to their resources (such as vocabulary knowledge, a little amount task specific data, the model size, etc.). Experimental results show that the vocabulary trim down algorithm made the model size being cut off 70% with almost the same accuracy of general model. After adapted by 143 minutes task specific data 27% word error rate reduction can be achieved comparing with the retrained model (using original general purpose data plus all available task specific data) in our Farewell99 dialog system.

1. INTRODUCTION

One of the great challenges for practical speech recognition systems is how to utilize user's resources for improving the performance of acoustic model. It is particularly useful in real applications to provide the user with the flexibility of dynamically configuring the size of the speech system model by the vocabulary of the given task. In most cases, a general-purpose acoustic model is trained from a large population's voice with balanced phoneme designation. However when this model is applied to given task directly, there is a need to adjust/adapt the model according to application's requirements such as task vocabulary, model size, etc. Furthermore, if users can provide some task specific speech data, we need to adapt the general model so that it can produce better performance in the given application.

The hierarchy of decision tree provides a good downsizing architecture [1]. Leaves of the same ancestor node represent HMM states of similar acoustic realization. Since the general acoustic model can be trained as many parameters as possible. Based on the general model's decision tree structure, all the parameters under some pre-chosen ancestor node are clustered down to a smaller size based on the criterion of minimizing the loss in the likelihood of generating the training data. Using the statistics embedded in the general acoustic model, the original training speech need not be referred in the downsizing algorithm.

One of the most exciting and promising areas of speech research is large vocabulary continuous speech recognition. A myriad of applications awaits a good recognizer. However, when so-trained general-purpose acoustic model is used in a specific task its recognition performance often degrades rapidly for there is a mismatch between the testing and the training conditions. It is impractical and inflexible of configuring a recognizer due to the tedious training process. Vocabulary-specific training is one of the mainly problem. With each new vocabulary comes the dilemma of tedious retraining for optimal performance, or tolerating substantially higher error rate [6]. In this paper, vocabulary based trim down algorithm and task adaptation is proposed attempting to alleviate vocabulary-specific training problem.

Firstly, we proposed a vocabulary-based trim down algorithm in decision tree-based acoustic model to make the model more close to the given task according to task specific vocabulary information. Then we do task adaptation based this model using interpolation method. Based this framework, users can configure the acoustic model by themselves according to their resources (such as vocabulary knowledge, a little amount task specific data, the model size, etc.). When we have a robust and accurate general model, we can easily tailor the model suitable for the given task.

2. VOCABULARY-BASED TRIM DOWN

Mixture Gaussian distributions and context dependent phone model have been used to achieve high performance in many continuous speech recognition systems based on continuous density hidden Markov models (HMMs)[2]. In this case, data insufficiency problem occurs owing to the increased number of phones and parameters. Furthermore, the distribution of phones and contexts is usually uneven in general speech and speech database. This fact requires a method to balance between model complexity and available training data. Decision tree state clustering based acoustic modeling has become increasingly popular for modeling speech spectra variations in large vocabulary speech recognition.

In decision tree state clustering based acoustic modeling, each node of the decision tree is attached to a question regarding to the phonetic context of the triphone units. A set of states can be recursively partitioned into subsets according to the phonetic questions at each tree node when traversing the tree from the root to its leaves. States reaching the same leaf node on the decision tree are regarded as similar and tied.

At first, we use the statistics knowledge embedded in the general-purpose acoustic model with decision tree based structure to downsize the model according to the vocabulary of the given task without using the original speech training data

[†] This work was done while the first author was working at Intel China Research Center.

that is usually very enormous. Given a general decision tree-based hidden Markov model that balances the model complexity and the training data efficiently, whereas when we decide to use the model in a specific task we expect that the model size can be tailored easily to match our resources. In our algorithm one of simple tailor principle is that if one state does not occurred in the given vocabulary then this state will be merged with its brother state. Furthermore, we tailor the decision tree in a high level. If any parent node with incomplete descendants it will be replaced by its child node that have complete descendants.

Given two leaf nodes n_1 and n_2, with Gaussian $G_1 = N(m_1, \Sigma_1)$ and $G_2 = N(m_2, \Sigma_2)$ respectively.

From Baum-Welch ML estimate:

$$m_1 = \sum_{x \in X} g(x)x \Big/ a \qquad (1)$$

$$\Sigma_1 = \sum_{x \in X} g(x)(x - m_1)(x - m_1)^T \Big/ a \qquad (2)$$

where X={speech data aligned to Gaussian G_1 with occupancy count $g(x)$ for each data x}, $a = \sum_{x \in X} g(x)$ is total occupancy of Gaussian G_1 in the training data. Assume that both sets of data X and Y are modeled by the combined Gaussian $G = N(m, \Sigma)$, i.e., when the two leaf nodes n_1 and n_2 are merged together. Refer to [1], the merged mean and variance can be computed as follows:

$$m = \frac{a}{a+b} m_1 + \frac{b}{a+b} m_2 \qquad (1)$$

$$\Sigma = \frac{a}{a+b}\left\{\Sigma_1 + (m_1 - m)(m_1 - m)^T\right\} + \frac{a}{a+b}\left\{\Sigma_2 + (m_2 - m)(m_2 - m)^T\right\} \qquad (2)$$

3. TASK ADAPTATION

It is well-known that automatic speech recognition (ASR) systems that have been designed for broad use perform poorly during a special application compared to systems that have been designed specifically for this very purpose. The reason for this degradation can be found in a mismatch between speaker characteristics, transmission channels, and task of the training data to those of the field assignment.

The recognition task can be mainly characterized by the vocabulary [3,4] and thus we need dealing particularly with the implications caused by changing dictionaries. The vocabulary-based trim down algorithm has been given in the above section that makes the acoustic model more compact and more closely to the specific task. However, this does not improve the performance in the given task directly.

Since a little more task specific data can be collected easily, if users can provide some amount of task specific data, the task adaptation component can make the general model (or the vocabulary based trim down model) adapt to the given task. Then, using the vocabulary trim down model as the initial model and the adaptation data as training data to obtain a task dependent model. At last, we do interpolation between the vocabulary-based trim down model (i.e., task independent model) and the task dependent model. This method is also called approximate maximum a posterior (AMAP) estimation scheme [6](see Fig.1).

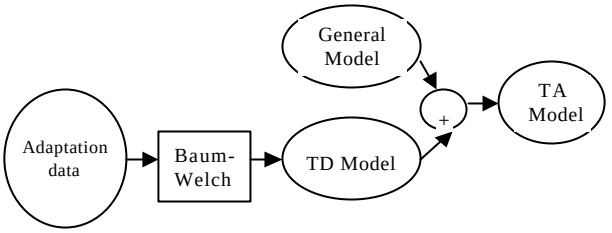

Fig. 1. HMM adaptation using AMAP method

Suppose that $g_t(s_t)$ is the probability of being at state s_t at time t given the current HMM parameters, and $f_{it}(s_t)$ is the posterior probability of the ith mixture component

$$f_{it}(s_t) = p(w_i \mid x_t, s_t) = \frac{p(w_i \mid x_t) N(x_t; m_i, \Sigma_i)}{\sum_{j=1}^{M} p(w_j \mid x_t) N(x_t; m_j, m_j)}$$

An approximate MAP (AMAP) estimation scheme can be implemented by linearly combining the general purpose and the task specific counts for each component density

$$\langle x \rangle_i^{AMAP} = l \langle x \rangle_i^{SI} + (1 - l) \langle x \rangle_i^{SD}$$

$$\left\langle xx^T \right\rangle_i^{AMAP} = l \left\langle xx^T \right\rangle_i^{SI} + (1 - l) \left\langle xx^T \right\rangle_i^{SD}$$

$$n_i^{AMAP} = l n_i^{SI} + (1 - l) n_i^{SD}$$

Where the superscripts on the right-hand side denote the data over which the following statistics (counts) are collected during one iteration of the forward-backward algorithm

$$\langle x \rangle_i = \sum_{t, s_t} g_t(s_t) f_{it}(s_t) x_t$$

$$\left\langle xx^T \right\rangle_i = \sum_{t, s_t} g_t(s_t) f_{it}(s_t) x_t x_t^T$$

$$n_i = \sum_{t, s_t} g_t(s_t) f_{it}(s_t)$$

The weight l controls the adaptation rate. Using the combined counts, we can compute the AMAP estimates of the means and covariances of each Gaussian component density from

$$m_i^{AMAP} = \frac{\langle x \rangle_i^{AMAP}}{n_i^{AMAP}}$$

$$\Sigma_i^{AMAP} = \frac{\left\langle xx^T \right\rangle_i^{AMAP}}{n_i^{AMAP}} - m_i^{AMAP}\left(m_i^{AMAP}\right)^T$$

4. EXPERIMENTS

The proposed approach of the trim down algorithm was evaluated on our Farewell99 dialogue system. The general model was trained from our telephony data as dictation acoustic model. 12 mel-cepstral plus their first and second order time derivatives were used as acoustic features. All HMMs have three emitting states and a left-to-right topology. Phonetic decision tree state tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones. The final triphone HMMs were built based on the tied states from the clustering. The number of mixtures for each tied state is 12 and the totally state number is 6k.

To our Farewell99 dialogue system, decoding was done using a graph decoder. The purely test for trim down algorithm is evaluated on telephony dictation test set, which consists of 110 utterances of 5 female and 6 male. The evaluation on trim down and task adaptation is performed on Farewell99's test set, which consists of 797 utterances of 3 female and 4 male.

Firstly, we have implemented the trim down algorithm proposed in [1]. The method of minimizing the loss of likelihood when merging the nodes of decision tree we call it the general trim down algorithm while the new vocabulary based trim down algorithm is named as vocabulary trim down. Table 1 gives the performance of two trim down method in the dictation test set and the Farewell99 test set. From table 1 we can see that the vocabulary trim down algorithm can down size the general purpose 6k model to relatively small size (1.7k states) with the almost the same performance of general trim down model with 3k states.

Table 1. Performance of two trim down algorithm

	Dictation	Farewell99
General purpose Model(6k)	14.4%	8.9%
General trim down(3k)	16.0%	9.7%
Vocabulary Trim down(1.7k)	16.5%	9.3%
General purpose Model(3k)	14.8%	9.8%

Then, we do task adaptation using vocabulary trim down 1.7k model and general trim down 3k model as seed model respectively. The results of task adaptation are shown in the fig.2. From 7 minutes task specific data to 406 minutes are being tested gradually. To be compared with retrained model using all dictation data and all task specific data (here is 406 minutes), the word error rate (WER) of the retrained model is also labeled in the figure. From the figure we can see that both adaptation scheme achieved good error reduction. After adapted by 66 minutes task specific data, the task adaptation (TA) models are almost the same good as the retrained model. However, do task adaptation is more convenient and more timesaving than retrain the model using all data directly. While adapted by 143 minutes data, the TA model obtained from vocabulary trim down algorithm outperforms the retrained model 27%. At last, when all available task specific data is used to adapt the vocabulary trim down model, the gain is 40% error reduction. From the figure we can see also that during adaptation the vocabulary trim down model always outperforms the general trim down model especially when the adaptation data is of little amount. We can think this is because the another merit of small size of model for the former model has only nearly half parameters as that of general trim down 3k model.

The weight l is the parameter that controls the adaptation rate. How to choose the appropriate value of l is a critical point while doing task adaptation. Usually, the value of weight is set up according to relative ratio between the statistics count of seed model's training data and that of task dependent data. Experiment results show that small value of l such as 0.02 or 0.01 can accelerate the adaptation rate so that they can contribute to good adaptation effect. Small value of the weight l means more emphasis on task dependent model during the interpolation. Another question is to use the same value of l in all states or each state has its own weight. The experiment results on these two adaptation rate weight selecting strategy are given in the fig.3. From the figure we can see that the state varying weight outperforms the constant weight when using less adaptation data. This is because when there is only little amount adaptation data, more states possess small occupancy so that if we accelerate these states' adaptation rate will contribute to good effort (only those states with occupancy greater than a pre-defined value can they be adapted in our training framework).

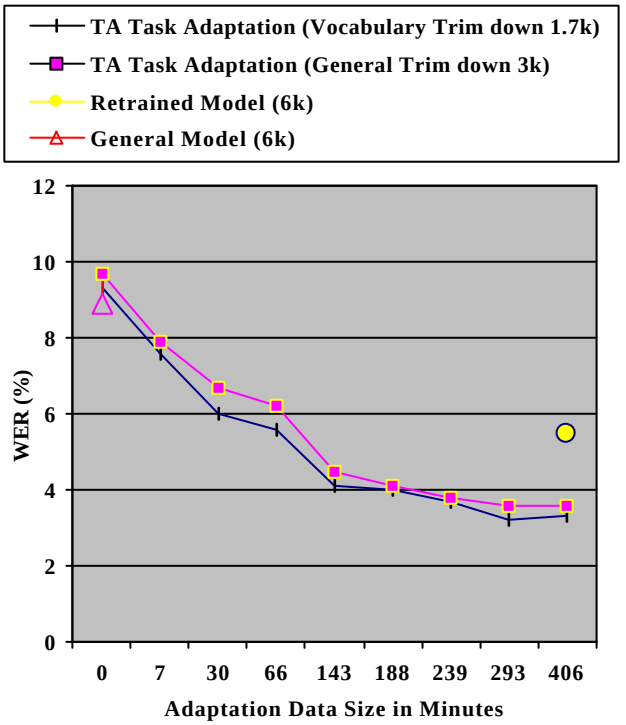

Fig. 2. Task adaptation based on vocavulary trim down model and general trim down model.

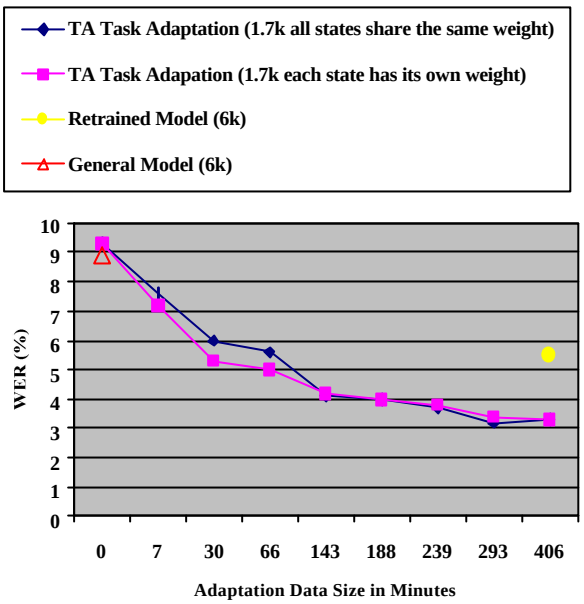

Fig. 3. Performance of two interpolating weight selecting strategy.

5. CONCLUSION

One of the great challenges for practical speech recognition systems is how to utilize user's resources for improving the performance of acoustic model. It is particularly useful in real applications to provide the user with the flexibility of dynamically configuring the size of the speech system model by the vocabulary of the given task.

In this paper, we proposed a vocabulary-based trim down algorithm in decision tree-based acoustic model to make the model more close to the given task according to task specific vocabulary information. Then we do task adaptation based this model using interpolation method. Based on this framework, vocabulary-specific training problem can be alleviated and users can configure the acoustic model by themselves according to their resources (such as vocabulary knowledge, a little amount task specific data, the model size, etc.) while not need to retrain the model via a tedious training process.

To our Farewell99 dialogue system, we use the method to tailor the acoustic model and do task adaptation. The experiment results show that the vocabulary trim down algorithm made the model size being cut off 70% with almost the same accuracy of general model. With 143 minutes adaptation data, the adapted trim down model contributes 27% word error rate reduction than the retrained model.

6. REFERENCES

[1] M.Y. Huang and X.D. Huang. Dynamically Configurable Acoustic Models for Speech Recognition. ICASSP98, 669-672.
[2] C. H. Lee, F. K. Soong and K. K. Paliwai. Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, 1996.
[3] V. Digalakis and L. Neumeyer. Speaker Adaptation Using Combined Transformation and Bayesian Methods. IEEE Trans. Speech Audio Processing, vol. 4, pp. 294-300, July. 1996
[4] Udo Bub. Task Adaptation for Dialogues via Telephone Lines. ICSLP96, 825-828.
[5] V. Digalakis, R. Rtischev and L. Neumeyer. Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures. IEEE Trans. Speech Audio Processing, vol. 3, pp. 357-365, Sep. 1995
[6] H. W. Hon, K. F. Lee. On Vocabulary-Independent Speech Modeling. ICASSP90, 725-728.