International Conference on Chinese Computing, pp. 106-110, Nov. 27-29, 2001, Singapore

The Definition and Extension of the Question Set for Decision Tree Based State Tying in Chinese Speech Recognition

Jing Li, Fang Zheng, Jiyong Zhang, Mingxing Xu, and Wenhu Wu
Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084
[lijing, fzheng, zjy, xumx, wuwh]@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn

Keywords: **Chinese speech recognition, Decision Tree, Question Set**

1. Abstract

This study deals with the decision tree based state tying method for acoustic modeling in Chinese Speech Recognition. In this paper, the definition of the context dependent Initial-Final units is given, and the linguistic knowledge based question set used in decision tree is described. The basic question set used in our experiment is based on the classified contexts. Two methods for extending and refining the basic the question set are also proposed in this paper. One is adding simple questions (corresponding to the unclassified contexts) to particular states of the units after investigating the influence of the contexts for the states. The other one is further adding two-side questions to the extended question set. In this way, the left and the right contexts are considered at the same time during the node's splitting. The experimental results show that the two methods can improve the performance of the acoustic model.

2. Introduction

Nowadays, the decision tree based state tying method is widely used to refine the acoustic model in large vocabulary speech recognition [1][2][3]. This approach is based on some prior knowledge of the language and the maximum likelihood optimization principle. In comparison with the bottom up data driven clustering approach, it has some advantages. Firstly, the model of any possible speech recognition unit can be estimated, even though that unit is not seen in the training data. The models of the unseen units are synthesized by the decision tree constructed in the training procedure. Secondly, it is convenient to adjust the number of the clusters to solve the contradiction of the robustness of the acoustic model and the sparseness of the training data.

However, there are some issues in decision tree based state tying method to be considered. For example, the single Gaussian mixture is used when a set of states are clustered and tied, and this may introduce some errors at the very beginning of the training procedure, which is harmful to the acoustic model. Some researchers are trying to find more precise methods to overcome this shortage [4]. Another very important issue is how to define an appropriate question set used in decision tree. And there are some researchers focusing on the question set refinement [5].

In our experiments, the single Gaussian mixture is still used when the states are tied, and our research concentrates on the extension and refinement of the question set in Chinese speech recognition, including:

- Design of the basic question set (BQS).
  Commonly, the question set is designed according to the context information. The left and the right contexts are classified into some clusters based on the linguistic knowledge. The question set is consisting of these questions. It is beneficial to the unseen or rarely seen units to use this question set.
- Investigating the influences of the questions for different units and states.
  This study may be helpful to selecting proper questions to extend and refine the question set.
- Extension and refinement of the question set.
  Two methods are used to extend and refine the question set. By adding simple questions to BQS, the extended question set 1 (XQS1) is obtained. Furthermore, two-side questions, which consider both the left and the right contexts, are added to XQS1 and hence the extended question set 2 (XQS2) is generated.

In this paper, the definition of the speech recognition units are given in Section 2, and the tree based state tying method is described in Section 3, the design and the use of the decision tree are presented in Section 4, the experimental results and some conclusions are given in the last two sections.

3. Definition of Speech Recognition Units

It is well known that the Initial-Final structure is a characteristic of Chinese syllable. The initial part corresponds to a consonant, and the final part corresponds to a vowel. Hence a lot of linguistic knowledge can be used to design the question set. There are 27 Initial units and 38 Final units used in this paper, as showed in Table 1.

Table 1: Definition of context-independent Initials/Finals

Initial units (27)	Final units (38)
b, p, m, f, d, t, n, l, *g, k, h, j, q, x,* *zh, ch, sh, z, c, s, r,* *_a, _o, _e, _I, _u, _v*	*a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn*

In Table 1, the Initial units "*_a*", "*_o*", "*_e*", "*_I*", "*_u*", "*_v*" are Zero-Initials. Some experimental results show that the model using the Zero-Initials is relatively better than that without using it. On the other hand, the number of possible context dependent units can be significantly reduced when Zero-Initials are introduced.

In our experiments, all these Initial/Final units are treated as the central units for the context dependent modeling. And after considering the left and right contexts of the central unit, the context dependent Initial/Final units called Tri-IF units are generated. In addition, the silence model is considered as a context-independent unit for garbage modeling.

4. Decision Tree Based State Tying

A phonetic decision tree is a binary tree in which a yes or no phonetic question is attached to each node. At the beginning, a set of states are gathered at the root node of the tree. And then a node splitting procedure will continue iteratively. If the states contained in an instantaneous (not the final) leaf node are not similar, all states in this node will be divided into two subsets according to the question to be asked, accordingly two new son nodes (new instantaneous leaf nodes) will be generated, corresponding to a yes answer and a no answer respectively. This procedure stops if a certain criterion is met. Therefore all states in the same leaf node are regarded to be similar and hence tied together. The structure of the decision tree is showed in Figure 1. The symbol like "*L_Stop?*" in the figure is the question attached to the node, and every leaf node corresponds to a certain model cluster.

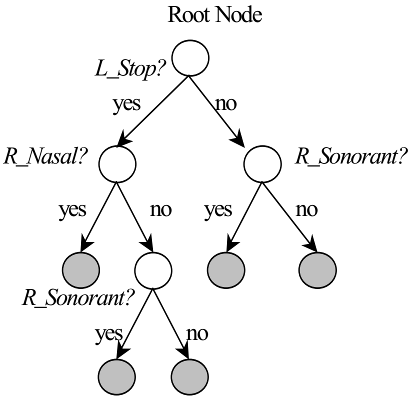

Fig 1. Decision Tree Structure

It is possible to re-estimate the log likelihood of the training data given any set of states. And this can be done without referring to the training data itself, since for single Gaussian distributions the means, variances and state occupation counts form sufficient statistics. During the splitting of tree node, the log likelihood will increase since it provides twice as many parameters to model the same amount of data. The best question that gives the biggest log likelihood improvement is selected as the node question. There are two thresholds used to terminate the split process, one is the total occupation count in a node, and the other one is the log likelihood increase. The split procedure stops, when the occupation count or the likelihood increase falls bellow the predefined thresholds.

5. Extending and Refining the Question Set

5.1.Definition of the BQS

The question set plays an important role in decision tree based state tying. An appropriate question set must be provided. The question set definition used in our experiments is based on linguistic knowledge [6], e.g. the classified left and/or right contexts. The following are some of the questions for the Final units:

QS_L_Stop:	*{b-*, d-*, g-*, p-*, t-*, k-*}*
QS_L_Nasal:	*{m-*, n-*, l-*}*
……	
QS_R_Stop:	*{*+b, *+ d, *+ g, *+ p, *+ t, *+ k}*
QS_R_Nasal:	*{*+m, *+n, *+ l}*
……	

And here are some of the questions for the Initial units:

QS_L_OpenN:	*{an-*, en-*, ang-*, eng-*}*
QS_L_HighFront:	*{i-*, u-*, v-*}*
……	
QS_R_OpenN:	*{*+an, *+en, *+ang, *+eng}*
QS_R_HighFront:	*{*+i, *+u, *+v}*

……

The question "*QS_L_Stop*" means that "*Is the left context of the current unit a Stop consonant?*" or "*Is the left context of the current unit in {b, d, g, p, t, k}?*". And the question "*QS_R_HighFront*" means that "*Is the right context of the current unit a high front tongue vowel?*" or "*Is the right context of the current unit in {i, u, v}?*". And these clustered contexts are used as the questions in decision tree based state tying.

Because every Chinese syllable consists of an Initial unit and a Final unit in our definition (with Zero-Initial defined), both the left context and the right context of an Initial unit must be either a silence or a Final and similarly both the left context and the right context of a Final unit must be either a silence or an Initial.

In the next two sub-sections, two methods to extend and refine the question set will be described. The first way is to add some simple questions to the BQS resulting in XQS1. The second is to add two-side questions to XQS1 to further extend the question set and hence XQS2 is obtained. A two-side question is one considering both the left and the right contexts of the central unit.

5.2.Adding simple questions to BQS for particular states

During the splitting of tree node, the best questions are chosen. In the maximum likelihood sense, the best question is the one that can achieve the maximal log likelihood when being asked. Our experimental results show that the influences of the left/right questions are different either for different units or for different states of the same unit during the root node's splitting. In this experiment, the BQS is used. Some results are shown in the Figure 2 and Figure 3.

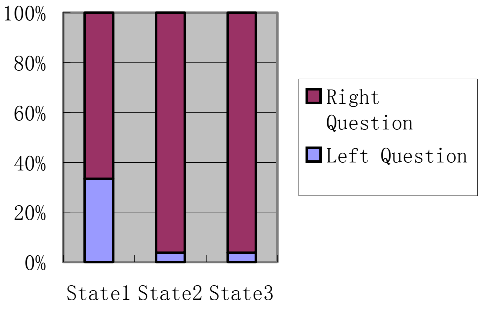

Fig 2. Proportion of the left/right questions chosen in the root node's splitting procedure for Initial units

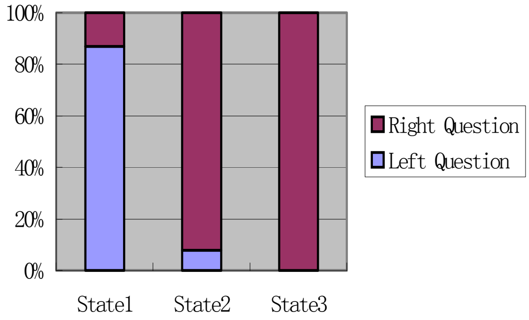

Fig 3. Proportion of the left/right questions chosen in the root node's splitting procedure for Final units

From Fig 2 and Fig 3 it can be seen that the influences of the left contexts and the right contexts are different for different states when the root node is being split. For example, there are about 66.67%, 96.30% and 96.30% of the best questions coming from the right question set for three states of Initial units respectively, during the root node's splitting. It means that the right contexts are more important than the left contexts for Initial units. And for the Final units, there are about 86.84% of the best questions coming from the right question set for the first state, and there are 92.11% and 100% of the best questions selected from the right question set for the second and third state, respectively. It means that for Final units, it is preferred to choose the right questions to split the first state, and on the other contrary, the left contexts to split the second and the third states.

The results show that for different states, the effect of the left contexts and the right contexts are different. So, it is necessary to introduce some simple questions to the question set in order to describe the variance of the states in detail. Some simple questions are as follows.

QS_L_b:	{b-*}
QS_L_p:	{p-*}
……	
QS_R_b:	{*+b}
QS_R_p:	{*+p}
……	

In our experiments, the left and right simple questions are added to the question set for tying the first state of the Initial/Final unit. And for the other states, only the right simple questions are considered, for the right contexts are more important. The recognition results and its comparison are shown in Section 5.

5.3.Extending and refining XQS1 by adding two-side questions

Another way to extend and refine the question set is to add two-side questions to it. Although the best question is often corresponding to the contexts of the same side of the central unit during the root node's splitting, it does not mean that the contexts of the other side are not important for the node splitting. Indeed, it only means that the influences of the other side's context are relatively weak. In the traditional method, only the single-side questions are asked during the node's splitting, which ignores the effect of the other side's contexts during the current node's splitting, and this would result in that some better splitting schemes are lost. Based on this analysis, a new method to extend the question set is proposed that considers both the left and right contexts simultaneously.

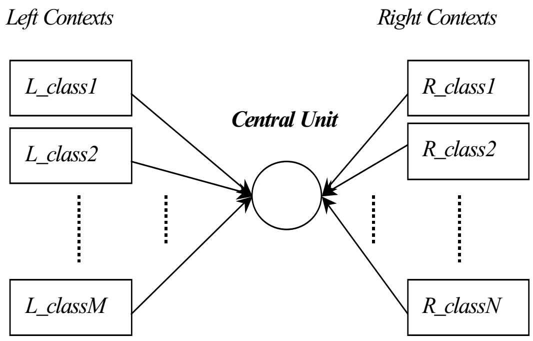

Fig 4. The classified contexts of the Initial/Final units

As shown in Figure 4, the left contexts are classified into some clusters {*L_class1, L_class2, ⋯, L_classM*} while the right contexts { *R_class1, R_class2, ⋯, R_classN* }. Notice that the same contexts can be included in several clusters. And then the two-side questions are constructed using these classified contexts by combining each left context and each right context. It will be like,

QS_L_class1_R_class1: {*L_class1-*+R_class1*}
QS_L_class1_R_class2: {*L_class1-*+R_class2*}
……
QS_L_classM_R_classN {*L_class1-*+R_classN*}

The question *QS_L_class1_R_class1* means that "*Does the left context belong to the L_class1 and the right context belong to the R_class1?*". These two-side questions are added to XQS1, and accordingly the new extended question set XQS2 is generated. The recognition results are shown in the next section.

6. Experimental Results

6.1.Speech database

The speech database used in our experiments is taken from the "863" male speech database [7], all the sentences of which are uttered in standard Chinese with a little regional accent with some background noise. There are 1,560 sentences, divided into three groups. Each speaker uttered one group of sentences. Totally, 80 males' speech data is used in our experiments. 70 males' data is used as the training set, and the other 10 males' data is used as the Testing Set 1. An additional testing set consists of 240 sentences, which was ever used for "the 863 Assessment" in 1998, referred to as Testing Set 2 in this paper.

6.2.Feature extraction

The features used in this paper are 34-dimensional mel-frequency cepstrum coefficients (MFCCs) [8], consisting of 17-demensional cepstrum as well as its first order derivatives.

6.3.Training

In our experiments, each unit is modeled using a left-to-right non-skip 3-state continuous density HMM. There are two major steps in the training procedure. Firstly, a single mixture context-independent (CI) Initial/Final acoustic model is trained. Secondly, a state tying context-dependent (CD) acoustic model is created. HTK v2.2 [9] is used in our experiments. The procedure is as follows in detail:

l Single mixture CI Initial/Final acoustic model is trained and then re-estimated using the Baum-Welch algorithm [10].

l The un-tied Tri-IF model is obtained by simply cloning its corresponding CI Initial/Final model. The new model is very big, and only the Tri-IF units seen in the training data are modeled. The parameters of the new model are re-estimated iteratively for several times using the same training data.

l Decision tree based state tying is then performed. Afterwards, a tied Tri-IF model is formed. It is necessary for the model to be re-estimated for several times. Three kinds of question sets are used in our experiments, the question set derived from the clustered contexts, i.e. BQS, as well as two extended question sets, namely XQS1 and XQS2, as mentioned above.

l Use the mixture split method to increase the number of the mixtures to refine the Tri-IF model.

6.4.Experimental results

The experimental results are shown in Table 2 and Table 3.

Table 2: Syllable accuracy over the 10-male testing set

Testing Set 1	1-mixture	2-mixture	4-mixture	Error rate reduction
BQS	65.93	69.61	72.77	*Baseline*
XQS1	69.50	72.15	74.31	5.7%
XQS2	70.36	72.92	74.93	7.9%

Table 3: Syllable accuracy over the 240-sentence testing set

Testing Set 2	1-mixture	2-mixture	4-mixture	Error rate reduction
BQS	71.03	74.49	77.68	*Baseline*
XQS1	75.68	78.03	80.16	11.1%
XQS2	75.65	78.35	80.99	14.8%

Table 2 shows the recognition results (syllable accuracy) over the 10-male testing set, and Table 3 shows the results over the 240-sentence testing set. The total state number is about 7,400 for the three question sets after decision tree based state tying is performed.

7. Discussion

We compare one basic question set based on the linguistic knowledge, and its two extended question sets, they are BQS, XQS1 and XQS2, respectively. The recognition results show that the syllable accuracy when using XQS1 can be significantly increased than that when using BQS both for the 10-male testing set and the 240-sentence testing set. The syllable error rate reduction is 5.7% for the 10-male testing set and 11.1% for the 240-sentence testing set. When XQS2 is used, the syllable error rate reduction can be as big as 7.9% for the 10-male testing set and 14.8% for the 240-sentence testing set. The use of XQS2 improves the performance better than the use of XQS1.

Some conclusions are as follows.

l The proposed methods to extend and refine the question set are efficient to improve the performance of the acoustic model.

l Indeed, these two methods are conflictive. The first method is based on that the influences of the left and right contexts are quite different; the use of single-side questions only emphasizes the quite important side's context, so the description of this side is refined. And the main idea of the second method is that the one side context information is still useful even it is not so important as the other side context in case that the influences of the two-side contexts are comparable to each other. These two methods aim at different situation. So the effects of these two methods are dependent on the training data.

8. Acknowledgements

We would like to thank Dr. Zhanjiang Song, Dr. Lei He, and Dr. Fan Wang for some helpful discussions on this research.

9. References

[1] Bahl, L. R., de Souza, P.V., Copalakrishnan, P. S., Nahamoo, D. and Picheny, M. A., "Decision trees for phonological rules in continuous speech", in Proc. Int. Conf. Acoustics, Speech, Signal Processing'91, Toronto, ON, Canada, May 1991, pp.185-188 (1991)

[2] Reichl, W. and Chou, W., "Decision trees state tying based on segmental clustering for acoustic modeling", in Proc. Int. Conf. Acoustics, Speech, Signal Processing'98, pp.801-804 (1998)

[3] Reichl, W. and Chou, W., "Robust decision tree state tying for continuous speech recognition", IEEE Trans. Speech and Audio Proc., 8(5): 555-566 (2000)

[4] Duchateau, J., Demuynck, K. and Van Compernolle, D., "A novel node splitting criterion in decision tree construction for semi-continuous HMMs", in Proc. Eurospeech'97, Rhodes, Greece, pp.1183-1186 (1997)

[5] Beulen, K. and Ney, H., "Automatic question generation for decision tree based state tying", in Proc. Int. Conf. Acoustics, Speech, Signal Processing'98, Seattle, WA, May 1998, pp.805-808 (1998)

[6] Wu, Z. J., "Outline of experimental phonetics", Higher Education Press, Beijing, (1989)

[7] Zheng, F., Song, Z.-J., and Xu, M.-X., "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine", EuroSpeech'99, Vol.2, pp.819-822, Budapest, Hungary (1999)

[8] Zheng, F. and Zhang, G.-L. "Integrating the energy information into MFCC", in Proc. Int. Conf. on Spoken Language Processing, vol.1, pp.389-292 (2000)

[9] Yong, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., The HTK Book (for HTK Version 2.2), Cambridge University (1999)

[10] Rabiner, L. R. and Juang, B. H., "An introduction to hidden Markov models", IEEE ASSP Mag., vol.3, no.1, pp.4-16, Jan. (1986)