CLASS-TRIPHONE ACOUSTIC MODELING BASED ON DECISION TREE FOR MANDARIN CONTINUOUS SPEECH RECOGNITION

GAO Sheng, XU Bo and HUANG Tai-Yi
National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy of Sciences
P.O.Box 2728, Beijing China, 100080
E-mail: {gsh,xubo,huang}@prldec3.ia.ac.cn

ABSTRACT

Decision tree based acoustic modeling has increasingly become popular for modeling speech spectral variations in continuous speech. In this paper, class-triphone acoustic models based on the decision tree are investigated for mandarin speaker-independent continuous speech recognition. Three main questions are discussed: how to select base phone models, how to generate the question set based on linguistics knowledge and how to produce class-triphone models through triphone-merging technique. To shorten the experiment time, extracting subtree algorithm is proposed and the number of the class-triphone models may be flexibly adjusted. The experimental results show that higher performance is obtained with class-triphone models than diphone models.

1. INTRODUCTION

The motive to build more robust acoustic models spurs the research on the variability of acoustic representations that occurs in the continuous speech. Much of the variability inherent in speech is due to the contextual effects. This means that a pronunciation of a phone is heavily dependent on its preceding and following phones[4]. By taking these contextual effects into account, the variability can be reduced and the accuracy of the models increased.

So much attention is paid to HMM models dependent on the context. As mentioned above, a phone is influenced by its left and right contexts. If we consider the left and the right at the same time when building models (model considering the left and right context, called triphone), the number of models and computation expeditiously increase. Then the problem, data-sparsity when training HMM models becomes austere. Especially while we consider the co-articulation in the inter-word, decoding is more complex. But if we only consider the left or right contexts, the number of acoustic models called diphone models, and computation decrease. So diphone models are first exploited. Experiments show that diphone models are more robust than context-independent models[6].

But in recent years, the research on acoustic models moved from diphone to triphone. For triphone models consider both the left and right contexts, they are more robust and preciser than diphone models. They have the advantage that they can predict the unseen triphones, which contexts do not occurred in training corpus. Through sharing model parameters and clustering similar models, the number of triphones and estimated model parameters can be reduced. When sharing and clustering, the top-down or bottom-up method is often used[3]. While constructing triphone models, decision tree based acoustic modeling has increasingly become popular for modeling the speech spectral variations in continuous speech recognition and better performance is achieved[1,2,3]. The process of generating decision tree is a data-driven one, and linguistics knowledge can be flexibly integrated into. For decision tree is controlled by many parameters, we can adjust these and optimized decision tree to obtain high performance.

But in mandarin speech recognition, diphone models are still popular and little investigation on triphone models is done. Chinese language has the monosyllabic structure in which each syllable consists of two parts, the initial part(声母) and the final part(韵母). This syllable structure is simple compared with the structure of English language. So the initial and the final is generally as the HMM model units. The previous work in our group[4] investigates diphones where the co-articulation

in the intra-syllable and inter-syllable is considered and its experiments shows that higher recognition accuracy with diphones is got and that the initial/final parts are influenced by both left and right context. In order to catch subtle variability of phones, we must simultaneously consider the left and right context and build triphones. Due to the unique structure of Chinese syllable, we must consider the inter-syllable contexts, but for English language the cross-word contexts may be ignored and only the intra-word contexts considered to make decoding simple. Decoding problem may be a reason that much attention is paid to triphone models.

In this paper, we build class-triphone models based on decision tree to improve the robustness of acoustic models. The experiments are done on mandarin large vocabulary continuous speaker-independent recognition system and compared with diphone models, higher recognition accuracy is achieved.

In the next section, we describe the base phones and how to construct the decision tree. In section 3, we describe how to generate triphones and how to merge triphones with the same output distributions. In section 4, some experimental results are showed. In section 5, the conclusion is drawn.

2. CONSTRUCTING DECISION TREE

While creating class-triphones, the most important step is to construct optimal decision tree. Decision tree is mainly influenced by four factors: how to select base phones, how to design the question sets based on linguistic knowledge, how to set the stop criterion and how to select evaluation function. Through adjusting these factors optimal decision tree and high performance is obtained. In the following we describe how to solve the above factors.

2.1 Base Phone Set

As mentioned in section 1, Chinese syllable has a monosyllabic structure, which consists of the initial part and the final part. The initial set consists of all consonants except “ng” and the null initial, which is the initial part of the syllables that begin with vowels. The finial set consists of all vowels including compound vowels and nasal-final which is the combination of vowels (or diphthong) with nasal-ending. So there are 22 initials and 37 finals altogether.

According to knowledge of Chinese monosyllabic structure, we design a base phone set denoted with P^1, which contains 22 initials, 37 finals and one silence. For consonants may be classified according to the first vowel phoneme of their following finals, that is /a/, /o/, /e/, /I/, /u/ and /ü/, there are 53 detailed initials. For example, /b/ has 3 detailed consonants, /b/, /bi/ and /bu/. Since detailed consonants consider the first vowel phoneme of the following finals, some finals are merged. For example, /an/ and /ian/ may be merged into /an/. Then the syllable /xian/ may be represented by /xi/ and /an/, not by /x/ and /ian/. There are 23 finals altogether. Then another base phone set denoted with P^2 is defined, which contains 53 initials, 23 finials and one silence.

In the paper we design the above two base phone sets according to our previous knowledge. To compare the performance of class-triphone models based on these sets, the experiment is made. The result shows that the recognition accuracy is approximately same, but the recognition error due to interpolation and deletion is more using P^2 base phone set than using P^1 base phone set. This may be due to discrimination capability of the final HMM models, for the number of the final class-triphones is much smaller in P^2 than in P^1 and the number of the initial class-triphones inversely. Although P^2 has a advantage when considering the syllable tones where the number of tonal class-triphones may decrease due to the small finals, we finally select the P^1 base phone set.

2.2 Question Set

The question set directly influences decision tree, for node splitting of the tree is controlled by the questions. The main rule is phone similarity, such as the similar manner of articulation and the similar phoneme. The question set consists of two parts, the left question subset and the right question subset. The left and right questions are mostly symmetric because there are no clear reasons for supposing that the preceding contextual factors would be different from the following ones. When we design the questions, the combination restriction of the initials and finals is considered, which reduces the number of question. For example, if the base phone is an initial, its right phones must be the finals and left phones must be the finals or the silence. Inversely if the base phone is a final, its right phones must be the initials or the

silence and the left phones must be the initials. So we respectively aim at the type of the base phone(initial or final) when building the question set.

In the question set each question related contexts contains some initials, finals or one silence, which have the linguistic or acoustic similarity, called a context set. The context set, which is comprised of some initials, is based on the similarity of the manner of articulation and the phoneme similarity. For example, according to the manner of articulation of the initials such as *stop, fricative, affricate* etc, the questions may be such as the following:

Example 1: /b/, /p/, /d/, /t/, /g/ and /k/
Example 2: /f/, /h/, /x/, /s/ and /sh/

According to the initial phoneme, the following is some question examples
Example 1: /d/, /t/, /z/, /c/, /s/, /n/ and /l/
Example 2: /zh/, /ch/, /sh/ and /r/

And the context set which includes some finals is based on the phoneme similarity of the first vowel in the final if the question is a left one, or the phoneme similarity of the last vowel in the final if the question is a right one. We list some left and question examples about the finals.

Right questions:
Example 1: /a/, /ai/, /ao/, /an/ and /ang/
Example 2: /ü/, /üe/, /ün/ and /ün/
Left questions:
Example 1: /a/, /ia/ and /ua/
Example 2: /an/, /ian/, /uan/, /üan/, /en/, /in/, /un/ and /ün/

In our question set, we also consider the middle vowels in the compound vowels, such that /a/, /iang/ and /uang/ may be occur in a context set because they all have /a/. Other characters are that some initials or finals may occur in many context sets and that one context set may be a subset of another context set, for example:

Question 1: /b/, /d/, /g/, /p/, /t/ and /k/
Question 2: /b/, /d/ and /g/

So based on above method and according to Chinese linguistics knowledge, we adjust the question set which finally contains 78 questions, including 37 left questions and 41 right questions.

2.3 Stop Criterion

The stop criterion is to ensure that every leaf node in decision tree occupy enough samples in order to robustly re-estimate the parameters of HMM models. In the paper we set a minimal sample number contained in a leaf node as the threshold. If the sample number in a node is less than the threshold when splitting, mark it as a leaf node.

2.4 Evaluation Function

The evaluation function is to evaluate the sample similarity in a node of decision tree. It may be one of the distance measures, such as the mean square distance. Let L be the evaluation function and $X = \{X_1, X_2, \cdots, X_N\}$ be the total N samples contained in a node called parent node. Let $X^1 = \{X_1^1, X_2^1, \cdots, X_{N1}^1\}$, $X^2 = \{X_1^2, X_2^2, \cdots, X_{N2}^2\}$ be samples contained in two child nodes derived from the parent node and $X = X^1 \bigcup X^2$, $X^1 \bigcap X^2 = \Phi$. The value of the evaluation function on the parent node and the child nodes is respectively denoted by L_{parent}, L_{child}^1 and L_{child}^2. Let $\Delta = L_{child}^1 + L_{child}^2 - L_{parent}$ be the increase value.

In every node we select a question from the question set and calculate the increase value Δ when splitting the parent node into two child nodes according to the question. Then we select the question with the maximal increase value, split the parent node into two child nodes according to the question.

In the paper samples of each node is described by the probability density function.

2.5 Constructing Decision Tree

In our experiment, we use the sharing output distribution HMM models. Each output distribution of a base phone has a binary decision tree. The root node has a lot of features labeled with the question attributes, which describe their contexts. To get the labeled features, the 60 base phone models are trained using the standard Baum-Welch or viterbi algorithm. Then all training speeches are segmented into the output distributions

and labeled with their contexts.

When building decision tree, we start from the root node. In each node which sample number is more than the minimal value, we split the node into two nodes based on a question which gives the maximal increase value of the evaluation function, one *yes* node which answers yes to the question and another *No*. The detailed process is as same as [1].

To shorten the experiment time and optimize the decision tree, we first construct a detailed decision tree with a lower threshold, which means that the tree occupies more leaf nodes. If we keep sufficient information, a sub-decision tree with a new high threshold may be extracted from the detailed decision tree by traversing the tree, checking the sample number in every node, comparing with the new threshold, deleting the node with the sample number less than the new threshold, and reordering the tree. Based on the approach a new decision tree is rapidly created when increasing the threshold.

Because the number of the leaf nodes in decision tree may be reduced when increasing the threshold, the number of class-triphones also decreases. Using the above method, we can quickly get the relation between the performance of recognition systems and the number of class-triphones.

3. GENERATE CLASS-TRIPHONES

After constructing decision tree, all triphones can be generated through traversing the tree from the root node. Let $\lambda(A,B)$ be a base phone model, A denotes state transition probability matrix of λ and B output distribution sets. If the base phone model has N_λ output distributions, It has N_λ decision trees.

Let $l_i^{N_\lambda} = \{leaf | leaf \in ith\ decision\ tree\}, i = 1,2,\cdots,N_\lambda$ denote the leaf node set of the ith decision tree in λ. Let $\tilde{\lambda}(\lambda, p_L, p_R)$ denote a triphone with the base phone model λ, left phone p_L and right phone p_R. To produce the triphone we must decide the N_λ output distributions, which are the leaf nodes of the N_λ decision trees of λ. The triphone generation algorithm works as follows:

Step 1: Select a decision tree from N_λ decision trees

Step 2: Start from the root node of the selected tree.

Step 3: Traverse the selected tree from the root node. In each node, check the recorded question attributes. If the left phone p_L or the right one p_R is consistent with answered questions in the node, that means p_L or p_R occurs in any context sets of answered questions, then jump to the *yes* node of the current node. Otherwise jump to *no* node. If the node is a leaf node, then terminate and record the leaf node. The probability density function of the leaf node is an output distribution of $\tilde{\lambda}(\lambda, p_L, p_R)$.

Step 4: Go to step 1 until N_λ decision trees have been traversed.

When these N_λ output distributions are obtained, the triphone model $\tilde{\lambda}(\lambda, p_L, p_R)$ is determined.

As mentioned in section 2, many triphones do not occur due to some combination restrictions between the initials and finals. Although we consider these restrictions, there are about 26,340 triphones. If not merging, most triphones can not be robustly trained. Fortunately, due to decision tree many triphones with the same base phone have the same output distributions, which mean the same leaf nodes of the decision trees. We may merge some triphones and create a class-triphone. The contexts of the new class-triphone present all the contexts of the merged triphones. When all class-triphones are built, the standard Baum-Welch algorithm or viterbi algorithm is used to train these class-triphones. When the new-trained triphone models are obtained, we may re-segment and label training speeches. Then decision tree may be rebuilt to generate a new tree, or kept the tree structure invariable and re-estimate the parameters of tree nodes. The performance of these two approaches is approximately same according to our experiments. So in our experiment we use the latter.

4. THE EXPERIMENT RESULTS

We use the above approach to build decision tree and create class-triphones. The experiment results are compared with the ones of diphone models based on our mandarin speaker-independent continuous recognition system. The question set, the stop criterion and the number of triphone models have not been properly optimized. We respectively experiment with the male and female test speeches. Training speeches contains 10,925 continuous sentences by 20 speakers and 4266 words by 33 speakers. The male and female test speeches both contain 240 continuous sentences by 6 male speakers and 6 female speakers.

Decoding is only with acoustic models and without language model. The recognition syllable accuracy is listed in Table 1 and Table 2.

Table 1 Recognition syllable accuracy based on class-triphone models

Sex	Model Number	Output Distribution Number	Mixture Number	Syllable Accuracy (%)
Male	3782	2129	4	74.75
Female	4542	2646	4	75.29

Table 2 recognition syllable accuracy based on diphone models

Sex	Model Number	Output Distribution Number	Mixture Number	Syllable Accuracy (%)
Male	138	549	4	67.5
Female	138	549	4	68.9

The results show that class-triphones outperform diphones. The analysis to class-triphones indicates that many class-triphones are unseen in training database but occur in the test speech database. So class-triphones not only depict the seen contexts occurred in training database but also predict unseen class-triphones. This means that class-triphone models improve the robustness of HMM models.

The next experiment is to get the relation between the performance of recognition system and the number of class-triphones. We adjust the threshold of the stop criterion to produce different decision trees with different number of leaf nodes and to obtain different number of class-triphones. In the experiment we use the subtree-extracting method mentioned in section 2. The experiment result based on male database is showed in table 3.

Table 3 Recognition syllable accuracy based on different class-triphone models(mixture number=4)

Model Number	823	1864	2244
Syllable Accuracy (%)	71.33	73.09	74.50

This shows that the recognition syllable accuracy decreases when the model number is reduced. To keep the advantage of class-triphones we must keep more models than diphone. We may balance between complexity and computation of recognition system and the performance.

5. CONCLUSION

The next research is to optimize the question set, decision tree and class-triphones to obtain higher performance. For the number of class-triphones is much more than the number of diphone models, the search space increases rapidly, especially when integrating language model into decoding. Therefor how to reduce the search space is an exigent solving problem. This large search space makes the recognition system respond very slowly. We must investigate the algorithm which may reduce the search space and boost the recognition speed.

REFERENCES

[1] L.R Bahl, P.V. de Souza, P.S. Gopalakrishnan, D.Nahamoo, and M.A. Picheny, “Decision Tree for Phonological Rules in Continuous Speech”, ICASSP 89, Glasgow, May 1989, pp.185-188.

[2] W.Reichl and W.chou, “Decision Tree State Tying based on Segmental Clustering for Acoustic Modeling”, ICASSP 98, pp.801-804.

[3] Mei-Yuh Hwang, Xuedong Huang and Fileno A.Alleva, “Predicting Unseen Triphones with Senones”, IEEE Transactions on Speech and Audio Processing, Vol 4, No6, November 1998, pp.412-419.

[4] Bin Ma, Taiyi Huang, Bo Xu, Xijun Zhang, Fei Qu, “Context-Dependent Acoustic Models in Chinese Speech Language”, ICASSP’96, USA, May, 1996

[5] 林焘,王理嘉,<<语音学教程>>,北京大学出版社.

[6] 徐波,<<汉语非特定人听写机系统集成与研究>>,中科院自动化所博士论文.