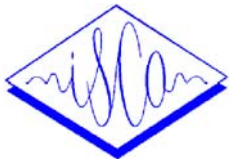

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DECISION TREE-BASED TRIPHONES ARE ROBUST AND PRACTICAL FOR MANDARIAN SPEECH RECOGNITION

LIU Yi and Pascale FUNG

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology, HKUST
Clear Water Bay, Hong Kong
{eelyx, pascale}@ee.ust.hk

10.21437/Eurospeech.1999-218

ABSTRACT

In large-vocabulary, speaker-independent speech recognition systems, modeling of vocabulary words by subword units is mandatory. This paper studies the use of triphone units for Mandarin speech recognition compared to biphone and context-independent phonetic units. In order to solve unseen triphones in speech recognition, decision-tree based clustering is used in triphone units. This method achieves high recognition performance with limited training data and also reduces the model training time. The robustness and effectiveness of the cross-word, tree-based triphone units have been proved by the speaker-independent continuous Mandarin speech recognition task. The training computation time reduces by about 2.3 times after tying states for triphone models, the recognition syllable accuracy increases 28.7% compared to monophone units and by 13.5% compared to biphone units.

1. INTRODUCTION

There is a surge in interest for research on LVCSR for Mandarin recently [1,2]. Many existing Mandarin recognition systems are built on Initial/Final units. However, the co-articulation effect between Initials and Finals is very strong in continuous speech [3], which leads to degradation in recognition accuracy. Thus, it is important and useful for us to study context-dependent units other than Initial/Final units. Among context-dependent units, triphone models consider both left and right phonemes at the same time, taking into account the co-articulation effects. Generally, triphone models are more robust than biphones and context-independent models. However, there are several practical problems when triphone models are applied in large-vocabulary speech recognition system. First, the model number is very large, the training and testing time becomes very long. Second, in large-vocabulary speech recognition, many triphones have only very few occurrences in the training data, hence there is no sufficient data for robust parameter estimation of these rarely seen triphones [4]. Finally, there are a large number of triphones missing in the training corpus. Unseen triphones are unavoidable because nearly all the combinations of phonemes are needed during decoding. In order to solve these problems, we propose using triphone units based on decision tree for acoustic modeling in Mandarin speech. Decision trees have been used to model allophones as a top-down generalization approach [5]. The number of triphone units and HMM models is reduced by tying or clustering similar models, so that the training time is decreased and the training data is more effectively used. Moreover, unseen triphones can be replaced by equivalent triphones.

The paper is organized as follows: Section 2 introduces the characteristics of Mandarin speech. In section 3, we describe how to generate the phonetic decision tree for triphone units in Mandarin. The experimental results comparing triphone units to biphones and monophones units are given in section 4. We conclude in section 5.

2. BASIC PHONETIC STRUCTURE OF MANDARIN

Mandarin is a tonal monosyllabic language. There are four lexical tones, and only 408 "base syllables" instead of 1345 "tonal syllables". Conventionally,

each syllable is divided into “Initial/Final” parts and the tone is associated with final part. This syllabic structure is simpler than western languages. The list for 21 initials and 37 finals used as basic units for triphone models is as follows:

Initials:
b, d, g, p, t, k, s, sh, x, f, h, z, zh, j, c, ch, q, n, m, k, l.

Finals:
a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, uen, ueng, ui, un, uo, uu, uuan, uue, uun.

In most cases, the Initial is very short compared to the Finial part. In continuous speech, the Initials are strongly affected by their preceding and following Finals. For continuous Mandarin speech recognition, triphone units can model this co-articulation effect better than other phonetic units. However, the number of triphone units used in Mandarin speech is over 20,000 without tying or clustering, which leads to long training time. The training data is also inadequate for all triphone units. Since some triphone units might not have enough training data, so we apply the decision tree to the triphone units.

3. CONSTRUCTION OF THE DECISION TREE

A phonetic decision tree is a binary tree in which a yes/no phonetic question is attached to each node [6]. Initially, all states in a given item list are placed at the root of the tree. Depending on each answer, the pool of states is successively split into smaller pools, and this continues until the states have propagated to leaf-nodes. All states in the same leaf node are then tied. Depending on the answer to the questions, all of the states end up in one of the shaded terminal nodes. Using decision tree for triphone units has three main advantages: First, it can lead to high quality state cluster for reducing training time. Second, the training data can be efficiently used for robust estimation. Third, unseen triphones can be synthesized by using triphone from similar states.

There are four important factors to be determined in a decision tree:

(1) The basic phones for triphone models
(2) The phonetic question associated with each node
(3) The evaluation function
(4) The stop criterion

3.1 Select the Basic Units

As mentioned in section 2, we use Initial/Final as basic phones for generating triphones. All initials and finals are context-independent, so the total number for Initials is 21 and 37 for Finals.

3.2 Phonetic Questions For Decision Tree

The splitting of the tree is controlled by phonetic questions. The questions selected are simple categorical questions, which are based on pronunciation gesture and linguistics knowledge [7,8]. We query about the left or right context of a triphone, such as “is the left or right context of the triphone a fricative?”. Generally, the phonetic questions are symmetric. Due to Chinese monosyllabic structure, the basic frame of triphone units is *Initial+Final+Initial(sil)* and *Final+Initial+Fina (sil)*.

For example, some of the questions concerning the pronunciation gesture of Initials and Finals are:

For Initials:

"Left_unvoiced-stops"{m_p-,m_t-*,m_k-}*
"Left_unvoiced-fricatives"{m_s-,m_sh-*,m_x-*,m_f-*,m_h-*}*
"Right_nasals" {+m_m,*+m_n}*
"Right_voiced-stops"
{+m_b,*+m_d, *+m_g}*

For Finals:

"Right_a-class"
{+m_a,*+m_ai,*+m_an,*+m_ao,*+m_ang}*
"Left_ang-class"
{m_ang-,m_iang-*,m_uang-*}*
"Left_ong-class" {m_ong-,m_iong-*}*

The phonetic question set for Mandarin we used consists of 49 questions, 26 for right questions and 23 for left questions.

3.3 State Tying

In practice, an unclustered triphone system based on single Gaussian continuous density output

distributions is built first. The log likelihood of the decision tree node on the training data can be derived from these single Gaussian distributions. The states are then grouped into the root node of the initial decision tree. This node is then split by using the phonetic question out of a question set which yields the biggest log likelihood improvement $\Delta(A,B)$ for the child nodes A and B [9]:

$$\Delta(A,B) = LL_{child}(A) + LL_{child}(B) - LL_{parent}(AB)$$
$$= -\frac{1}{2}\left(n_A \sum_{d=1}^{D} \log\left[\frac{\sigma_{d,AB}}{\sigma_{d,A}}\right]^2 + n_B \sum_{d=1}^{D} \log\left[\frac{\sigma_{d,AB}}{\sigma_{d,B}}\right]^2 \right)$$

Where n_X is the number of observations for node X, D the dimensionality of the feature vector and $\sigma_{d,X}$ the variance of component of node X.

3.4 Stop Criterion

Useful criteria for decision tree includes minimal delta entropy and minimal occurrence counts of each node [5]. Since we use top-down structure decision tree for Mandarin triphone units, the occupation count for the leaf node is selected as the threshold for the stop criterion. The threshold is useful in preventing the creation of clusters with very little training data. Similarly, any splits which would result in a total occupation count falling below the threshold is prohibited.

4. EXPERIMENTAL RESULTS

We use cross-word tree-based triphone units generated from the above method for speaker-independent continuous Mandarin speech recognition, and use HKU93 Mandarin speech database in our experiments. The database includes speaker-dependent isolated syllables and continuous speech in read style from a total of 20 speakers (10 males and 10 females). The database is designed to include all Mandarin syllables in all tones. We use speech data from 18 speakers for training, (it includes over 15,000 continuous sentences and all the syllables), 2 speaker's speech data (1 male and 1 female) for testing. The acoustic features used in the experiments are 13 MFCCs, 13 Delta MFCCs and 13 Acceleration MFCCs, 5 states HMM used for acoustic modeling, Gaussian mixture number is 4. The number of unclustered triphone units for the training data is 26340.

Table.1 shows the computation time for training. (The unclustered triphone units*, tied states triphone units**, monophone and biphone units are all 1 Gaussian Mixture). From the results, we can see that the training time is reduced by almost 2.3 times compared to unclustered triphone models.

CPU time	Mono-phone	Bi phone	Tri-phone *	Tri-phone **
Hours	1.333	2.333	8.833	2.667

Table.1 Computation time for training

Table.2 shows the recognition results of tied-state triphone models, biphones and monophones. We separately select 350 continuous sentences from the two testing speakers for testing data. Moreover, 350 continuous sentences including the male and female testing data are also used for testing. As expected our method shows higher accuracy than using monohones and biphones.

In Fig.1, "m_t-m_ua+m_d" is an unseen triphone, no training data for this triphone unit is found. But when it traverses the decision tree which we established in Fig.1, it can be modeled by the seen triphones instead of backing-off to biphones or monophones models. The circle in Fig.1 means a set of clustered triphone units.

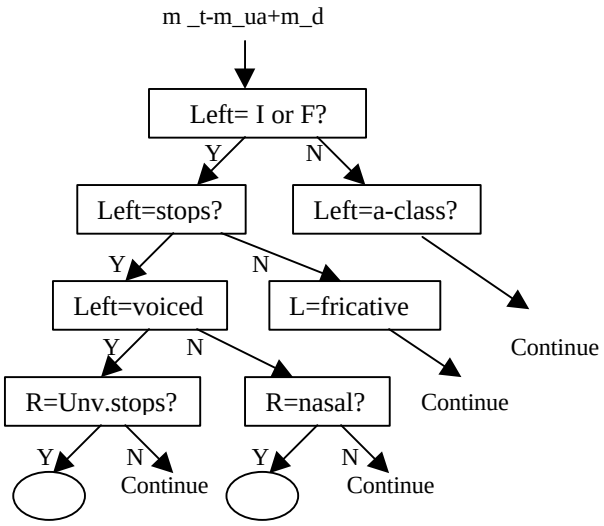

Fig.1 The decision tree for unseen triphones

Data Class	Monophone	Biphone	Triphone	Improvement	
	Syllable Acc.	Syllable Acc.	Syllable Acc.	For Monophone	For Biphone
Male	60.29%	66.67%	70.06%	16.20%	3.39%
Female	61.03%	67.97%	74.68%	22.34%	9.87%
Male & Female	54.48%	61.77%	70.12%	28.70%	13.5%

Table.2 The syllable accuracy of three units

5. CONCLUSION

In this paper, decision tree based triphone units is used for speaker-independent large vocabulary Mandarin speech recognition. The training time is reduced by 2.3 times after tying states for triphone units compared to untied triphone units, and is close to biphone training time. The syllable accuracy for Mandarin increases from 54.48% with monophone units, and from 61.77% with biphones units, to 70.12% with triphones units. Moreover, all unseen triphones in decoding are handled by the decision tree. In our future research, we will focus on designing better questions for Mandarin speech and automatic generation the linguistic questions for multi-lingual systems.

6. ACKNOWLEDGMENT

This work is partly supported by the Hong Kong Government's Central Allocation Grant CA97/98.EG02.

7. REFERENCE

[1] Lin-shan LEE. "Voice Dictation of Mandarin Chinese," IEEE Trans. Signal Processing, Vol.14, No.4 pp63-101

[2] Stephen W.K.Fu, C.H.Lee et al., "A Survey on Chinese Speech Recognition"

[3] L.Villarrubia et al., "Context-dependent units for vocabulary-independent Spanish speech recognition," ICASSP96 pp451-454

[4] W.Reichl and W.Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," ICASSP98 pp. 801-804

[5] M.Y. Hwang et al., "Predicting Unseen Triphones with Senones," IEEE Trans. Speech and Audio Processing, Vol.4, No.6 pp. 412-419 Nov.1996

[6] Steve Yong et al., "The HTK Book for HTK Version 2.1,"

[7] LIU Ming "Mandarin Book1" 1984, Chinese University of HK

[8] Gao Sheng, Huang Taiyi et al., "Class-triphone acoustic modeling based on decision tree for Mandarin continuous speech recognition," ISCSLP98 pp.44-48

[9] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," ICASSP98 pp. 805-808