Decision Tree Clustering for Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition

Po-yu Liang[2], *Jia-lin Shen*[1], *Lin-shan Lee*[1, 2]
1. Institute of Information Science, Academia Sinica
2. Department of Electrical Engineering, National Taiwan University
Taipei
E-mail : jlshen@iis.sinica.edu.tw

ABSTRACT

This paper presents the decision tree clustering for parameters tying in acoustic modeling for speaker-independent Mandarin telephone speech recognition. The triphone models considering both syllable internal variability and cross syllable variability are first investigated to better estimate the contextual acoustics and co-articulation in Mandarin telephone speech recognition. In order to predict unseen triphones as well as increase the trainability of the speech models, the decision tree based tying algorithms are used to share those distributions with similar linguistic characteristics. Experimental results show that the decision tree clustering approach leads to a 6.42% of error rate reduction.

1. INTRODUCTION

The context dependent acoustic models have been widely used in automatic speech recognition, i.e., each speech model represents a phone with specific left and right contexts. There are inherent problems with the context dependent acoustic models: the trainability and the unseen units, due to the insufficiency of the training data to cover all the context variations well for each context-dependent speech unit. In this study, a total of 34 context-independent (CI) phone-like units (PLUs) were first considered for Mandarin Chinese as shown in Table 1(a), but this number is immediately increased to 4605 after including both the left and right context dependencies, i.e., constructing triphones [1-2]. Not only the number of units is increased by more than 135 times but also very often more than half of the units do not occur in a normal speech database. There are several ways to increase the trainability using the triphone models, such as back-off and sharing methods [3-5]. Decision tree based tying algorithms are one kind of sharing technique, which were developed to solve the trainability as well as the unseen triphone problems for the western language [4-5]. The decision trees are constructed starting from a single root node representing all contexts. As each node is created, an optimal question is selected from a finite question set by maximizing some pre-defined criterion, e.g., likelihood increase or entropy decrease. In this study, such decision tree based tying algorithms were applied to Mandarin Chinese, compared to other acoustic modeling approaches. There are four elements in constructing the decision tree: question set, root node selection, split criterion and convergence condition. The first question set was transformed from a set used for English [5], including a total of 132 questions for left and right contexts, respectively. Different root nodes for generating the decision trees were also compared in this study, including the context independent (CI) and right context dependent (RCD) phone sets as well as the model and state levels, as will be clear later on. The entropy-like criterion was used to reduce the uncertainty of the decision tree when splitting a node. The number of total states in the tied triphone set were used as the convergence condition. Furthermore, a second question set obtained from Mandarin phonological and phonetic analysis [6-7] was also studied as a comparison.

The remainder of this paper is organized into 3 sections. Section 2 describes the decision tree based tying algorithm. The experimental results are discussed in section 3. Section 4 finally gives the concluding remarks.

2. DECISION TREE CLUSTERING

The decision tree based tying algorithm is a linguistic-driven top-down clustering method. Figures 1 and 2 show the clustering examples for

the phoneme "b(+u)". There are four elements in constructing the decision tree: question set, root node selection, split criterion and convergence condition. They are described in detail as follows.

2.1. The Question Set

As mentioned above, two question sets are used in this study. The first question set was transformed from a set used for English [5]. A total of 132 questions were obtained, including 10 general questions, 16 vowel questions and 40 consonant questions for left and right contexts, respectively. Considering the language differences between English and Mandarin Chinese, another set obtained from Mandarin phonological and phonetic analysis [6-7] was also studied as a comparison. The second one has 112 questions consisted of 11 general questions, 14 vowel questions and 31 consonant questions for left and right contexts, respectively. Figures 1 and 2 show a comparative example using the decision tree clustering for the phoneme "b(+u)" based on these two question sets, separately. One can find that different clustering results can be obtained based on different question sets. For example, in the first question set, the nasals "N" and "M" are always in the same cluster while in the second question set, they are allowed to be separated to different classes. Therefore, as shown in Figure 2, the nasals "N" and "M" can be clustered into different classes.

2.2. Root Node Selection

The root node selection of the decision tree determines the number of trees, which can be dependent of contexts and distribution levels. In context-independent (CI) phone model based decision tree clustering, 34 trees for the 34 PLUs as shown in Table 1.(a) can be obtained, whereas the number of trees are immediately increased to 34*N if every state of the phone model (with N states) is used as the root node, separately. Also, in RCD phone model based decision tree clustering, the decision tree based tying algorithms are applied to the left states only, which indicates a total of 480 decision trees for the 480 between-syllable RCD PLUs. This is because the between-syllable RCD phone models provide the best recognition performance among all the baseline experiments as shown in Table 2, which will be discussed later on. So we intend to apply the decision tree clustering based on the most successful between-syllable RCD PLUs.

2.3 Split Criterion

The entropy-like criterion is used to choose the optimal one among all the questions to maximize the increase of likelihood scores or the reduction of the uncertainty in splitting a node. Suppose the Gaussian distributions are used to represent the observation features in a node, the measure of entropy decrease (likelihood increase) ΔL can be expressed as in the following [8] :

$$\Delta L = \log|\Sigma| - \frac{n_1}{n_1+n_2}\log|\Sigma_1| - \frac{n_2}{n_1+n_2}\log|\Sigma_2|, \quad (1)$$

where Σ means the covariance matrix of the parent node while Σ_1 and Σ_2 denote the covariance matrices of the corresponding child nodes with n_1 and n_2 observation feature vectors, respectively.

2.4. Convergence Condition

To end up the growth of the decision trees, some convergence conditions must be met. Here the minimum likelihood increase on each splitting and the minimum number of observation vectors in each node are used in the first stage. They are empirically selected to maintain the sensitivity and trainability of the speech models. Furthermore, after all of the decision trees are constructed, the amount of leaves can be pruned to a pre-defined number using the bottom-up merging criterion based on minimum entropy decrease. In this way, we can compare the experimental results using the same number of states based on different question sets, root nodes, splitting and convergence criteria.

3. EXPERIMENTAL RESULTS

3.1 Speech Database and Initial Processing

The speech database was produced by 59 male and 54 female speakers over the telephone provided by Telecommunication Laboratory in Taiwan. Each speaker produced 120 Mandarin sentences such that a total of 13,560 Mandarin sentences (5.87 hrs) are included in the speech database. In the following experiments, 51 male and 49 female speakers were used to train the gender-dependent, speaker-independent models and the rest 8 male and 5 female speakers were used as the testing speakers. In the feature extraction process, after end-point detection is performed, 32 ms Hamming window is applied every 10 ms with a pre-emphasis factor of 0.95. 14-

order mel-frequency cepstral coefficients (MFCC) were derived from the power spectrum filtered by a set of 30 triangular band-pass filters. In addition, the first order derivatives of the 14 mel-frequency cepstral coefficients as well as the first and second order derivatives of the log short-time energy were also calculated to result in a feature vector of 30 dimensions for each frame. The left-to-right continuous hidden Markov model (CHMM) was trained for each speech unit and the number of mixtures per state is dynamically determined by the amount of training data with a maximum of 8 mixture components [2]. In addition, the cepstral mean subtraction (CMS) technique was used as the front-end robust processing.

3.2 Experiments

In Mandarin speech recognition, the most widely used units are the 22 Initials and 40 Finals, where Initial means the initial consonant and Final means the vowel part but including possible media and nasal ending [9]. This is because of the mono-syllabic structure of the Mandarin Chinese, in which each Mandarin syllable can be decomposed into an Initial/Final format. Table 1(b) shows the corresponding phoneme sequences for the Initial/Finals. However, if both the right and left context dependencies are included, the numbers for Initial/Final will be increased to 13,336. Note that the amount of Initial/Final units is nearly 3 times of that of triphones (4605) considering both the left and the right contextual variations. In the first experiment shown in Table 2, we compare the recognition results for different types of phone based and Initial/Final based acoustic models. One can find that comparable recognition rate for within-syllable right-context-dependent (RCD) Initial /Final and phone models can be obtained (50.74% vs. 49.19%). However, when the between-syllable right contextual effects are included, the phone based recognition results outperform that of Initial/Final by 3.10% (58.56% vs. 55.46%) with less than one half of mixture components as shown in Table 2 (7701 vs. 17015). It is noted that when the triphones are used, the recognition rates are even slightly degraded (56.80% vs. 58.56%) using more than 7 times of mixture components as compared to the between-syllable RCD phone models. This is probably due to the insufficient amount of training data.

Then the decision tree based tying algorithms are used to tie the parameters of the triphone models. As shown in Table 3, the root nodes are chosen based on CI phones and RCD phones, respectively, where a single decision tree generated for each phone or each state is also compared. The convergence condition is based on the total state numbers of the models. Note that the decision tree clustering approaches based on RCD PLUs outperform that based on CI PLUs. Also, better improvements can be achieved when the root nodes are changed from models to states, where around 4% of recognition rates increase can be obtained using half of mixture components both in CI phone based and in RCD phone based decision tree clustering approaches as shown in Table 3. It can be found that the best recognition results are achieved using the decision tree generated based on RCD phone units, in which a total of 2460 states are used. In comparison with the triphone models as listed in Table 2, the recognition rates are increased from 56.80% to 60.77% and the mixture numbers are significantly reduced from 55,272 to 10,638. The obtained recognition rates also outperform that using between-syllable RCD phone models (58.56% vs. 60.77%).

In the last experiment as shown in Table 4, the newly developed question set is used and the recognition accuracy can be further improved from 60.77% to 61.22% based on same configurations except for the question sets. As mentioned previously, the newly developed question sets come from Mandarin linguistic analysis such that the context variations for each phoneme can be grouped more accurately. It is believed that the recognition performance can be further improved with better question sets.

4. CONCLUSION

In his paper, we describe the improved acoustic modeling in speaker-independent continuous Mandarin speech recognition over the telephone using decision tree clustering. In order to predict unseen triphones as well as increase the trainability in triphone based speech models, the decision tree based tying algorithms are used. The question set, root node selection, split criterion and convergence condition in decision tree based tying algorithms are investigated and discussed. Experimental results show that improved performance can be obtained, which indicates an error rate reduction of 6.42%.

REFERENCES :

1. R.Y. Lyu, H.M. Wang & L.S. Lee, “A comparison of different units applied to isolated/continuous large vocabulary Mandarin speech recognition”, in *Proc. Int. Conf. Computer Processing of Oriental Language*, May 1994, pp. 211-214.
2. C.H. Lee & B.H. Juang, “A survey on automatic speech recognition with an illustrative example on continuous speech recognition of Mandarin”, *Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, Aug. 1996, pp. 1-36.
3. J.B. Marino, A. Nogueiras, A. Bonafonte, “The demiphones : an efficient subword units for continuous speech recognition”, *Eurospeech*, pp. 1215-1218, 1997.
4. M.Y. Hwang, X. Huang, F. Alleva, “Predicting unseen triphones with senones”, in *Proc. ICASSP,* Vol. 2, 1993, pp. 311-314.
5. J. J. Odell, “The use of context in large vocabulary speech recognition”, *Ph.D. dissertation*, Queen’s college, UK, Mar. 1995.
6. “國音學”, 國立台灣師範大學國音教材編輯委員會, Taiwan, 1982.
7. 王理嘉, “音系學基礎”, 語文出版社.
8. M.Y. Hwang, X. Huang, “Dynamically configurable acoustic models for speech recognition”, in *Proc. ICASSP*, Vol. 2, pp. 669-672, 1998.
9. L.S. Lee, "Voice dictation of Mandarin Chinese", *IEEE Signal Processing Magazine*, Vol. 14, No. 4, pp. 63-101, July 1997.

(a)

b	p	m	f	d	t	n	l	g	k	h	j	<	T	Z	C	S
R	z	c	s	a	o	e	E	i	u	U	r	M	N	#	Y	y

(b)

ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ
b	p	m	f	d	t	n	l	g	k	h	j	<	T	ZY	CY	SY	RY	zy	cy	sy
ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ	ㄧ	ㄨ	ㄩ	ㄧㄚ	ㄧㄝ	ㄧㄞ	ㄧㄠ	ㄧㄡ	ㄧㄢ
#a	#o	#e	#E	#ai	#Ei	#au	#ou	#aM	#M	#aN	#N	#i	#u	#U	#ia	#iE	#iai	#iau	#iou	#iaM
ㄧㄣ	ㄧㄤ	ㄧㄥ	ㄧㄛ	ㄨㄚ	ㄨㄛ	ㄨㄞ	ㄨㄟ	ㄨㄢ	ㄨㄣ	ㄨㄤ	ㄨㄥ	ㄩㄝ	ㄩㄞ	ㄩㄣ	ㄩㄥ	ㄦ	NULL Initial	NULL Final 1	NULL Final 2	
#iM	#iaN	#iN	#io	#ua	#uo	#uai	#uEi	#uaM	#uM	#uaN	#ueN	#UE	#UaM	#UM	#ioN	#r	#	Y	y	

Table 1 : (a). 34 phone-like units (PLUs). (b). The corresponding phonetic alphabets with respect to 22 Initials/40 Finals of Mandarin Chinese in terms of 34 phonemes.

Model	Mixture number	Accuracy (%)
CI phone	408	31.74
Intra-LCD phone	1920	43.43
Intra-RCD phone	1740	49.19
Intra-demiphone	2448	50.39
triphone	55272	56.80
Inter-RCD phone	7701	58.56
Inter-demiphone	10480	57.04
CI Initial/Final	904	44.56
Intra-RCD Initial/Final	1948	50.74
Inter-RCD Initial/Final	17015	55.46

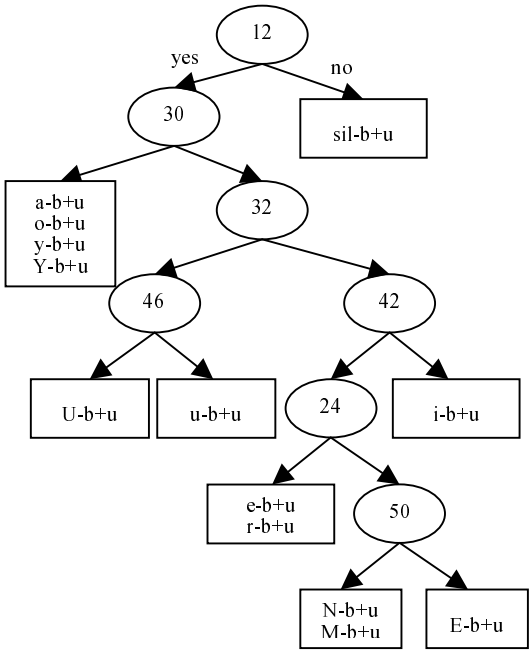
Table 2. The comparative results for different types of phone based and Initial/Final based speech units (intra- denotes within-syllable and inter- denotes between-syllable).

		State number	Mixture number	Accuracy (%)
CI phone based	model-level	1500	6000	56.10
	state level	3000	9020	58.84
		4500	16883	59.34
Inter-RCD phone based	model-level	4500	18000	56.21
	state level	1960(1000+960)	9052	60.20
		2460(1500+960)	10638	60.77
		2960(2000+960)	11628	60.42

Table 3. The experimental results based on decision tree tying algorithms.

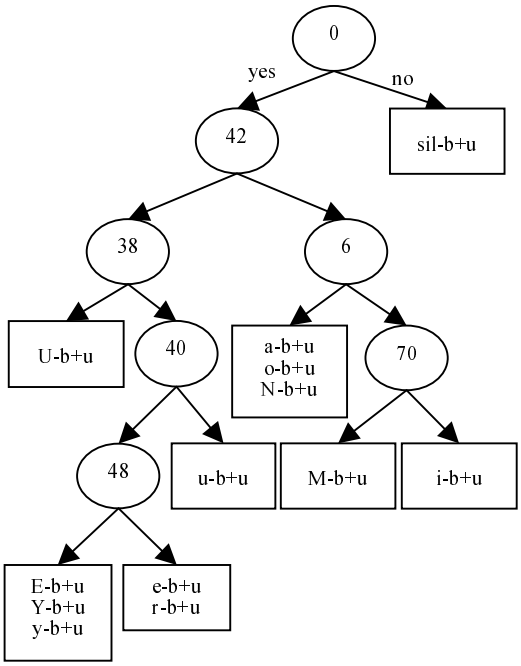
	Accuracy(%)
Question set 1	60.77
Question set 2	61.22

Table 4. The experimental results using decision tree based tying algorithm based on different question sets.



12:Is left context vowel?
24: Is left context back-vowel?
30: Is left context low-vowel?
32: Is left context rounded-vowel?
42: Is left context 齊口呼?
46: Is left context 撮口呼?
50:Is left context 尾音?

Figure 1 : A decision tree clustering example using the first question set for phoneme "b(+u)".



0:Is left context 具元音性?
6: Is left context 集聚性?
38: Is left context 撮口呼?
40: Is left context 開口呼?
42: Is left context 具高音性?
48:Is left context 具細音性?
70: Is left context 具柔潤性?

Figure 2 : A decision tree clustering example using the second question set for phoneme "b(+u)".