EuroSpeech, 3:1617-1620, Sept. 3-7, 2001, Aalborg, Denmark

Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition

Jiyong Zhang, Fang Zheng, Jing Li, Chunhua Luo, and Guoliang Zhang

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084
[zjy,fzheng,lij,luoch,zhanggl]@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn

Abstract

This paper describes the new framework of context-dependent (CD) Initial/Final (IF) acoustic modeling using the decision tree based state tying for continuous Chinese speech recognition. The Extended Initial/Final (XIF) set is chosen as the basic speech recognition unit (SRU) set according to the Chinese language characteristics, which outperforms the standard IF set. An adaptive mixture increasing strategy is applied when splitting the single Gaussian into mixed Gaussians in each tied state after the decision tree has been constructed. Our experimental results show that these two improvements are helpful to the acoustic modeling of Chinese speech recognition and that the CD XIF model outperforms the baseline syllable model over 30%.

1. Introduction

The acoustic modeling is one of the primary processes in large vocabulary continuous Chinese speech recognition systems. Unlike the western languages, Chinese is naturally a syllabic language and each syllable has an Initial-Final structure. Previous experiments show that the acoustic modeling based on either the syllable or the IF can achieve a good performance. Many Chinese speech recognition systems have been established, using syllables or IFs as the SRUs [1][2][3].

In systems where the context information is completely or partly not taken into account to reduce the computational complexity, the accuracy is often not satisfying because the co-articulation of the speech is not well modeled. In order to gain higher recognition accuracy, the CD modeling becomes an important part in the acoustic modeling. For Chinese speech recognition systems, two major issues should be considered. One is the selection of SRUs [4], while the other is the reduction of the scale of the CD model.

Typically there could be three types of basic SRUs for the Chinese CD acoustic modeling: syllables, IFs, or phones. Normally each syllable has a definite meaning and a stable pronunciation, but there are 418 toneless syllables totally. Choosing syllables as SRUs may result in totally about 70 million possible CD units. Therefore the computation and storage complexities would be unbearable for the time being. Compared with the syllable, the phone unit is rather small and there are only a small number of phones, but phones vary very much in pronunciation. There are often phone deletions, phone insertions and phone changes in continuous speech. However, IFs are relatively steady. There are only 59 standard Chinese IFs, which meet the computational complexity needs of the CD modeling. Based on the above analysis, the IF set could be a good compromise between the syllable set and the phone set, and hence the IFs are chosen as the SRUs for CD modeling in this paper.

To reduce the computation complexity in the CD modeling, the decision tree based state tying method is often adopted. The statistical framework of the decision tree provides two major advantages over the previous rule-based or bottom-up-based approaches [5]. First, the classification and prediction abilities of the decision tree allow the synthesizing of the model units or contexts, which do not occur in the training data. Second, being a model selection process, the node splitting procedure provides a way to maintain the balance between the model complexity and the model scale from a limited amount of training data.

In this paper, we build the CD IF models based on the decision tree with the state tying technique and make two further improvements. These improvements enhance the robustness and performance of our acoustic model.

In the experiments both the standard IF set and XIF set are studied, and the XIFs are selected as the basic SRUs. We consider the left and right co-articulation of the focused unit based on the XIF set to build the CD model. The decision tree based state tying process is performed as follows. Single-Gaussian models are trained first and the decision tree is constructed for each state of each central unit, and all the states with the same central unit are clustered using the decision tree based state tying algorithm. The resulting tied states are then retrained and the single-Gaussian HMM is estimated. Finally an adaptive mixture increasing strategy is adopted to split the single Gaussian into mixed-Gaussians for each tied state.

The structure of this paper is as follows. In Section 2 the two SRU sets are described. The detailed design of the CD IF modeling based on the decision tree is explored in Section 3. In the following section comes the adaptive mixture increasing strategy. The experimental conditions and results are given in section 5. And at last we come to the conclusion in Section 6.

2. Selection of the Basic Units

The selection of the basic SRUs is an important issue for Chinese speech recognition. In this section, we will first give the standard IF set, and then the XIF set definition.

2.1. The Standard Initial/Final Set

In the Chinese language there are 21 Initials and 38 Finals as in Table 1[6].

Table 1: The Standard Initial/Final (IF) Set

Type	Unit list
Initial (21)	*b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r*
Final (38)	*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*

2.2. The Extended Initial/Final Set

Mostly each Chinese syllable is consisting of an Initial and a Final, while some of them have only the Final part. The syllable without an Initial is called a Zero-Initial one. For example, the Zero-Initial syllable /*ang*/ only consists of Final "*ang*" while the syllable /*zhang*/ consists of Initial "*zh*" and Final "*ang*". Thus it is normal to see two adjacent Finals in a Chinese syllable string (sentence).

For the consistency purpose in the CD modeling so that the Initial and the Final will be seen one after another only, we extend the standard IF set by adding six Zero-Initials: "*_a*", "_o", "_e", "_i", "_u" and "*_v*". For example, the syllable /*ang*/ is regarded as consisting of Zero-Initial "*_a*" and Final "*ang*". The XIFs are shown in Table 2.

Table 2: The Extended Initial/Final (XIF) Set

Type	Unit list
Initial (27)	*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*
Final (38)	*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*

When we select the XIFs as the basic units, each Final is conjoint with two extended Initials at both sides; so two adjacent Finals will never be seen. This rule will make a great decrease of the total number of units in the CD modeling. For instance in our experiments, there are 122,118 possible units in the CD IF model while only 29,047 possible units in the CD XIF model.

3. Decision Tree Based State Tying

In the state tying technique, the design of the question set is one key point. In this section, we will explain how to design the question set for the state tying according to the Chinese linguistic knowledge and how to construct the decision tree, respectively.

3.1. Question Set Design

The point of the question set design is the similarity of pronunciation. As we choose the IFs as the basic SRUs, the selection of questions becomes rather simple and explicit. Considering each style of the pronunciation, we will obtain the left and the right questions. For example, two questions are corresponding to Type *Affricate*:

QS "R_Affricate" {*+z, *+zh, *+j, *+c, *+ch, *+q}
QS "L_Affricate" {z-*, zh-*, j-*, c-*, ch-*, q-*}

The token "+" represents the right co-articulation while "-" the left co-articulation. We can also consider the combined type, for example, Type *Aspirated* with Type *Affricate*:

QS "R_AspiratedAffricate" {*+z, *+zh, *+j}
QS "L_AspiratedAffricate" {z-*, zh-*, j-*}

The pronunciation style of the Final part can also be used to construct the questions. For example, the questions of judging whether the left and right context is vowel *"a"* could be designed as follows respectively:

QS "R_Type_A" {*+a, *+ai, *+an, *+ang, *+ao}
QS "L_Type_A" {a-*, ia-*, ua-*}

Note that the unit lists of the above two questions are not symmetric because the Final part has different influence on its left and right neighbor units. Based on the above method and the Chinese linguistic knowledge, we design a question set with 32 left questions and 29 right ones, totally 61.

The question set is related to the basic SRUs. When we select the XIFs as the basic SRUs, There are some modifications between the question set of the IF model and that of the XIF model. For example, in the question set of the XIF model, the question for judging whether the left context is vowel *"a"* could be modified as:

QS "L_Type_A" {*_a-*, a-*, ia-*, ua-**}

3.2. Constructing Decision Tree

Usually, a decision tree is built using a top-down sequential optimization procedure starting from the root node of the tree [7]. Initially, all the data with the same central unit are pooled together at the root node, which is also the only leaf node. Each leaf node is split according to the question that can result in a maximum likelihood increase on the training data. The process is repeated until the likelihood increase is smaller than a predefined threshold.

Let $X=\{x_1, ..., x_T\}$ be the sequence of observation vectors from the training data. The log-likelihood of the training data, $L(S)=\log P(X|S)$, generated from a tree node S can not be easily calculated, the common EM auxiliary function is used as the objective of the clustering [8]:

$$Q(S) = \sum_{x_t} \sum_{s \in S} \gamma_s(x_t) \log N(x_t \mid \mu(S), \Sigma(S)) \qquad (1)$$

where $\gamma_s(x_t)$ is the *a posteriori* probability of the observation x_t at Frame t generated from State s, and $N(\bullet|\mu,\Sigma)$ the Gaussian density with mean μ and covariance matrix Σ. To reduce the computation complexity, each node is described by a single Gaussian initially.

Because of the monotonic relation between the auxiliary function and the likelihood, the sequential optimization of the auxiliary function in the decision tree clustering also results in the optimization of the likelihood function. Therefore the auxiliary function can be used as objective in the decision tree. By using the single mixture Gaussian assumption for the cluster distribution, the likelihood variation of clustering can be efficiently evaluated for every tree node S. In each splitting process, tentatively each question is used to split the states

into two subsets S_{yes} and S_{no}. The question with a maximum increase for the auxiliary function is selected to split the node.

$$\Delta Q_q(S) = Q(S_{yes}) + Q(S_{no}) - Q(S) \quad (2)$$

When the maximum increase is smaller than a predefined threshold, the split process of the decision tree stops.

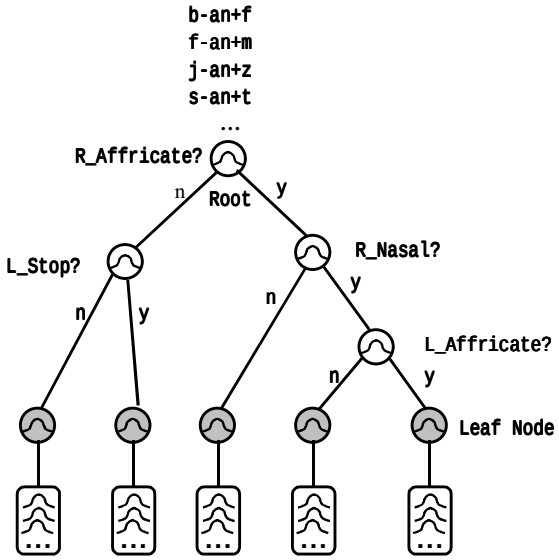

Figure 1.Decision Tree Based State Tying

Figure 1 shows the split process of decision tree for the central unit "*an*". The units with the same central unit construct the root node. And then units each node are divided into two subsets based on a question which gives the maximal increase value of auxiliary function $Q(s)$. During the split process, each node is described by a single Gaussian. When the leaf node is reached, a multiple-mixture Gaussian distribution will be adopted to replace the previous single Gaussian distribution to get higher accuracy.

4. Adaptive Mixture Increasing Strategy

In the decision tree building process, each node is only represented by a single Gaussian. To obtain a higher accuracy, we can increase the number of mixtures in each leaf node. A simplest way is to split the density in each leaf node into the same number of Gaussian mixtures. But since the training sample number for each tied state is unequal, this simple treatment may result in that some states are over trained while some ill trained.

As a solution, we propose the adaptive mixture increasing strategy, where the number of mixtures is determined according to the training data amount. The following function is used to describe the relationship between the mixture number $f(x)$ and the sample number x:

$$f(x) = \begin{cases} 1, & x \le 20 \\ \left\lceil \frac{x}{20} \right\rceil, & 20 < x \le 220 \\ 12, & 220 < x \end{cases} \quad (3)$$

Therefore the mixture number $f(x)$ is adaptive to the sample number.

5. Experimental Results

A continuous Chinese speech corpus from 863 materials is used [1]. The corpus contains 80 speakers' data and 520 utterances are available for each speaker. All the recorded materials are obtained in a low noise environment through a close-talk noise-canceling microphone. 70 speakers' data are used as the training set while the remaining part is used for testing. They are digitized at a sampling frequency of 16kHz. 13 dimensional MFCC, 1 dimensional log energy, and their 1st and 2nd order derivatives extracted in 24ms Hamming windowed frames every 12 ms are used as the features. Each unit is divided into 3 states, and the tools in HTK v2.2 are used for building our acoustic models [9]. The experimental results below are given in the acoustic level.

5.1. IF Set vs. XIF Set

In Table 3 the experimental results for the context-independent (CI) IF and XIF modeling are given.

Table 3: IF set vs. XIF set (CI modeling)

Number of mixtures	Syllable accuracy rate (%)	
	IF set	XIF set
1	43.71	47.09
2	50.13	55.26
4	54.25	58.67

From the above results, we can conclude that the XIF set outperforms the IF set with about a 4 absolute percent increase. Thus in the following experiments, the XIF set will be taken as the basic SRU set.

5.2. Constructing Decision Tree

Two more factors that may influence the construction process of a decision tree include the selection of the question set, which is usually done according to a prior knowledge, and the threshold definition is used to judge the clustering procedure of each SRU, which mainly depends on the training samples. The stop criteria with different threshold may cause dissimilar decision trees. And the selection of the threshold is an empirical value. Some experimental results for various thresholds are given in Table 4.

From this table, we can see that the threshold can be used to adjust the scale of the decision tree effectively. Both the total number of states and the syllable accuracy rate are increasing with the decrease of the threshold.

Table 4: Decision Tree with Different Thresholds

Threshold	Number of states	Syllable accuracy rate (%)
350	9,311	75.24
200	14,630	76.19
100	21,222	76.23

In our system, there are 42,255 states before the state tying process. When the threshold changes from 200 to 100, the syllable accuracy rate increases very slightly while the state number increases by about 45%. The decision tree with threshold=100 may make the states insufficiently tied which may reduce the robustness of the acoustic model. So a

suitable threshold should be chosen practically. In our experiments, 200 as the threshold is better than the other two.

5.3. Mixture Increasing Strategy

To improve the performance, the mixed Gaussian densities instead of a single Gaussian density are used to describe the probability distribution for each leaf node in the decision tree. Table 5 shows the performance as a function of the number of mixtures.

Table 5: Performance of Mixture Increasing

Number of mixtures	Number of Gaussians	Syllable accuracy rate (%)
1	14,630	76.09
2	29,260	78.29
4	58,520	80.26
6	87,780	80.93
8	117,040	81.11
Adaptive	54,836	80.85

We can see that about 20% accuracy increase can be obtained by increasing the number of mixtures. When the mixture number changes from 6 to 8, the syllable accuracy rate increases only about 0.2% absolutely, but the scale of the acoustic model increases by 33%. It could be a large computational burden. So we think that the performance almost reaches its maximum at the mixture number 6.

When we adopt the adaptive mixture increasing strategy, the syllable accuracy rate is near to that for 6 mixtures, while the model size is near to that for 4 mixtures. The adaptive mixture increasing strategy can cause a little degradation of the accuracy while reducing much of the model complexity. So we can conclude that the adaptive mixture increasing strategy is promising.

5.4. Comparison with Syllable Model

To illustrate the performance of the CD XIF Model, we give the CI syllable model [1] as the baseline. Each unit in the syllable model has 6 states and each state has 8 Gaussian mixtures.

Table 6: Performance of Syllable Model

Number of mixtures	Number of Gaussians	Syllable accuracy rate (%)
8	19,293	72.21

Table 6 shows that the CD XIF model with the adaptive mixture increasing strategy outperforms the syllable model with a 31% improvement of the accuracy.

6. Conclusions

In this paper, we build the context-dependent acoustic model based on the decision tree for continuous Chinese speech recognition. The XIF set is proposed to be the basic speech recognition unit set according to the Chinese language characteristics, and the model based on the XIF set outperforms the model based on the IF set with about a 4 absolute percent increase. When splitting the single Gaussian into mixed Gaussians in each tied state after the decision tree has been constructed, we apply the adaptive mixture increasing strategy and increase the syllable accuracy rate by about 20% while maintain the model size to a limited scale. Compared with the baseline syllable model, the CD XIF model can increase the syllable accuracy rate by over 30%.

In our future work, we will try to improve the model accuracy by utilize more information of the linguistic knowledge such as tone, prosody, and explore more efficient approach into the acoustic modeling process.

7. References

[1] 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

[2] Zhang, J.-Y., Zheng, F., Xu, M.-X. and Li, S.-Q., "Intra-Syllable Dependent Phonetic Modeling for Chinese Speech Recognition," International Symposium on Chinese Spoken Language Processing, pp. 73-76, Beijing, 2000

[3] Li, J., Zheng, F., and Wu, W.-H., "Context-Independent Chinese Initial-Final Acoustic Modeling," International Symposium on Chinese Spoken Language Processing, pp. 23-26, Beijing , 2000

[4] Zheng, F., Wu, W.-H., Fang, D.-T., "Speech recognition units in the Chinese dictation machines,", 4th National Conf. On Man-Machine Speech Comm. (NCMMSC-96), pp.32-35, Beijing, 1996

[5] Reichl, W. and Chou, W., "Robust Decision Tree State Tying for Continuous Speech Recognition", *IEEE Trans. Speech and Audio Proc*., Vol.8, No.5: 555-566, 2000.

[6] Wu, Z.-J. "The Chinese Phonetics in 'Man-Machine Dialogue'", Chinese Teaching In The World, vol4, pp3-20,1997 (In Chinese)

[7] Breiman, L., Friedman, J., Olshen, R.-A. and Stone, C.-J., Classification and Regression Trees, Belmout, CA: Wadsworth, 1984

[8] Dempster, A.-P., Laird, N.-M. and Rubin, D.-B., "Maximum likelihood from incomplete data via the EM algorithm", J.R.Statist. Soc., vol.39, pp 1-83, 1977

[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