International Symposium on Chinese Spoken Language Processing, pp. 73-76, Oct. 13-15, 2000, Beijing

INTRA-SYLLABLE DEPENDENT PHONETIC MODELING FOR CHINESE SPEECH RECOGNITION

ZHANG Jiyong, ZHENG Fang, XU Mingxing, LI Shuqing

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084

zjy@sp.cs.tsinghua.edu.cn

ABSTRACT

A novel acoustic modeling method for Chinese speech recognition based on Intra-Syllable Dependent Phone (ISDP) set is proposed and practiced. The ISDP set extends the traditional phone set based on the intra-syllable information of Chinese phonetic knowledge. The acoustic models based on ISDP set (ISDPMs) have the following features: One, they are suitable for the case of a rather small scale of training data. Two, this scheme is an integration form of tri-phone modeling and syllable modeling. The mixed Gaussian densities are used to describe the feature space of each ISDP and the Viterbi algorithm is adopted for decoding process. In addition, the ISDP-syllable search tree is designed and presented to reduce the decoding complexity. Our Experimental result shows that the ISDP modeling is more flexible and faster than Syllable Modeling meanwhile it causes no much deduction of the performance.

Keywords: Chinese speech recognition, intra-syllable dependent phone based models (ISDPMs), mixed Gaussian density, ISDP-syllable search tree

1. INTRODUCTION

In the automatic speech recognition (ASR) system using Hidden Markov Models (HMMs) [1], the choosing of the speech recognition unit (SRU) is an important issue [2]. Because many speech recognition systems for western languages are naturally phone based, so phones, di-phones or tri-phones are often chosen as the SRUs. Unlike the western languages, Chinese Speech is naturally a syllabic language. Prior experiments show that acoustic modeling based on syllable can reach good performance. Many Chinese speech recognition systems are established based on syllable, such as the EasyTalk system [3][4].

Chinese language has an initial-final structure and generally every consonant consists of one consonant phoneme while every vowel one or several vowel phonemes. Tri-phones can be chosen as SRUs for Chinese speech recognition (CSR) [5]. The tri-phone models have disadvantages because they are totally driven by training data where no Chinese language knowledge is considered. On the other hand, if we choose syllable as the SRU, the Chinese speech has over 400 syllables so that it is difficult to consider the context when training. The model scale of both tri-phone models and syllable models are huge, if the training corpus is not large enough, the accuracy will decrease greatly due to lack of training data.

Here we propose the Intra-Syllable Dependent Phone set and take intra-syllable dependent phones (ISDPs) as SRUs. The ISDP set extends the traditional phone set based on the Chinese speech knowledge and the method of articulatory. The acoustic models based on the ISDP set (ISDPMs) has following features: firstly, they has smaller SRUs, thus they are more suitable for the case of a rather small scale of training data than the triphone models or syllable models. Secondly, as the Chinese speech is naturally syllable composed, the intra-syllable context correlation is more important than the inter-syllable context correlation. The ISDPMs embedded the intra-syllable context and ignore the inter-syllable context. From the syllable model's view, the ISDPMs is trying to share the train data across each individual syllable. From the tri-phone model's view, the ISDPM is a compact form of tri-phone model.

The acoustic modeling method used here is based on the Gaussian mixture distribution. The mixed Gaussian densities are used to describe the feature space of each ISDP. We adopt the Viterbi [6] algorithm for decoding process. In addition, the IDSP-syllable search tree is designed and presented to reduce the decoding complexity.

This paper is organized as follows. In the next section, we introduce the gaussian mixture model. The ISDP set is illustrated in detail in the following section. In the section 4, we design the ISDP-syllable search tree, which is used in the Viterbi decoding process to reduce the acoustic searching consumption. We give out the experiments and their results in the next section. In the final section we give an outline of the future work.

2. THE GAUSSIAN MIXTURE MODEL

Researches and experiments on HMM distance measures have shown that the transition probability matrix plays a far less important role than the observation probability matrix does. The Gaussian Mixture Model (GMM) is the reductive form of traditional HMM.

Obtained the feature vector set $Z_k = \{\vec{z}_1, \vec{z}_2, \cdots, \vec{z}_T\}$ of a state of the SRU, we use the following probability density function (pdf) to describe the distribution of $\vec{z}_j$:

$$p(\vec{z}_j|\theta) = \sum_{i=1}^{M} \left(p_i p(\vec{z}_j|\theta_i) \right)$$

where $\vec{z}_j \in Z$ is a d dimensional cepstral vector, $j = 1,2,\cdots,T$,M is the number of components of the mixture model.

$$p(\vec{z}_j|\theta_i) = (2\pi)^{-d/2}|R_i|^{-1/2}\exp\left(-\frac{1}{2}(\vec{z}_j-\vec{\mu}_i)^T R_i^{-1}(\vec{z}_j-\vec{\mu}_i)\right),$$

it is the pdf of a single Gaussian Model, its parameter θ_i includes the mean vector $\vec{\mu}_i$ and covariance matrix R_i . p_i is the weight of each single Gaussian Model, and $\sum_{j=1}^{M} p_i = 1, p_i \ge 0.$ θ is the parameter of GMM, it includes p_i and θ_i , i= $1,2,\cdots,M$ 。

Given Z, in which the cepstral vectors are assumed to be independent, our goal is to determine the GMM parameter θ, so that we get the maximum value of $p(Z|\theta)$,where $p(Z|\theta) = \prod_{j=1}^{T} p(\vec{z}_j|\theta)$ and $p(Z|\theta)$ is differentiable with respect θ. We can obtain the following equation:

$$f = \nabla_\theta\left(\ln p(Z|\theta)\right) = 0$$

By solving the equation, an iterative procedure can be applied to estimate $\vec{\mu}_i$, R_i and $p(\omega_i)$, i= $1,2,\cdots,M$ .We take $\vec{\mu}_i$ as an example:

$$\begin{aligned} f &= \nabla_{\mu_i}\left(\ln p(Z|\theta)\right) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1}\nabla_{\mu_i}\left(\sum_{i=1}^{M} p_i p(\vec{z}_j|\theta_i)\right) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1} p_i \nabla_{\mu_i} p(\vec{z}_j|\theta_i) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1} p_i p(\vec{z}_j|\theta_i) R_i^{-1}(\vec{z}_j-\vec{\mu}_i) \\ &= 0 \end{aligned}$$

then, we can get

$$\vec{\mu}_i = \sum_{j=1}^{T}\left(P_{i,j}\cdot\vec{z}_j\right) \Big/ \sum_{j=1}^{T} P_{i,j},$$

where

$$P_{i,j} = p(\vec{z}_j|\theta)^{-1} p_i p(\vec{z}_j|\theta_i)$$

We can revise the formula above by changing $\vec{\mu}_i$ to $\hat{\vec{\mu}}_i$, in this way an iterative procedure is formed to estimate the parameter. Similarly, $\hat{p}_i$ and $\hat{R}_i$ can be obtained. The following formula are for solving the problem of GMSM parameter estimation, they are on the basis of EM (Expectation-Maximization) algorithm.

$$\hat{\vec{\mu}}_i = \sum_{j=1}^{T}\left(P_{i,j}\cdot\vec{z}_j\right) \Big/ \sum_{j=1}^{T} P_{i,j}$$

$$\hat{R}_i = \sum_{j=1}^{T}\left(P_{i,j}(\vec{z}_j-\hat{\vec{\mu}}_i)(\vec{z}_j-\hat{\vec{\mu}}_i)^T\right) \Big/ \sum_{j=1}^{T} P_{i,j}$$

$$\hat{p}_i = T^{-1}\cdot\sum_{j=1}^{T} P_{i,j}$$

3. INTRA-SYLLABLE DEPENDENT PHONE

In this section we explain how to obtain the ISDP set. First we establish the basic phone set, which can be viewed as uni-phone set. Then we give the ISDP set in which the intra-syllable context is taken into account.

3.1 Basic Phone Set

Our basic phone set is consisted by three types of phones: initial phones, inter phones and final phones. Each initial part of the Chinese speech is viewed as one phone, which we call initial phone. Each final part of Chinese speech is consisted by one or two inter phones and one final phone while the pronouncing time is long. The inter phones can be viewed as the middle part of the syllable and the final phones as the last part. Here we give the content of the basic phone set:

Table1. Basic phone set

Phone type	Basic phone list
Initial phones	b, p, m, f, d, t, n, l, g, k, h, j, q, x, z, c, s, zh, ch, sh, r, a, o, e, i, u, v
Inter Phones	A, E, O, I, U, V
Final phones	A_, E_, O_, I_, U_, V_, N, G

To illustrate how each syllable consisted with the basic phones, here we give some examples of syllable to base phones:

Table2. Example of syllable to basic phones

Syllable	Initial phone	Inter phone	Final phone
/a/	a	A	A_
/han/	h	A	N
/shuang/	sh	U, A	G

3.2 Intra-Syllable Dependent Phone Set

From the above discussion we know that each syllable has the following structure:

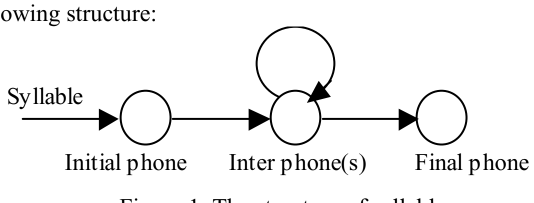

Figure 1. The structure of syllable

As the intra-syllable dependent phones ignore the inter syllable context, we only consider the right context of initial phones, left and right context of inter phones and left context of final phones. Here we do two reductive treatments. One reduction is that we ignore the right context of inter phones due to the relationship between inter phones and final phones has been embodied in the left context of final phones. Thus we can only consider the right context of initial phones and the left contexts of inter and final phones. The other reduction treatment is that we divide the initial phone into several classes by the method of articulatory, and we use the initial phone class instead of initial phone when we consider the left context of inter phones.

The initial phones are divided into the following classes by the method of articulatory:

Table3. The initial phone classes:

Class No.	Method of Articulatory	Initial Phones
0	Stops	b, d, g
1	Aspirated Stops	p, t, k
2	Affricates	z, zh, j
3	Aspirated affricates	c, ch, q
4	Nasals and Laterals	m, n, l
5	Fricatives	f, s, sh, r, x, h
6	a – type	a
7	o – type	o
8	e – type	E
9	i – type	i
10	u – type	u
11	v – type	v

There are total 150 ISDPs existed in the Chinese speech. Here we give some syllable examples:

Table4. Example of syllable to ISDPs

Syllable	Combination of ISDPs
/a/	aA + 6A + A_
/han/	hA + 5A + AN
/shuang/	shU + 5U + 10A + AG

4. THE ISDP-SYLLABLE SEARCH TREE

Unlike the Syllable Modeling, the recognition results of ISDP Modeling are ISDPs, we should convert this ISDPs to Syllables. Here the ISDP-syllable search tree is designed to reflect the relations among ISDPs and syllables so that the acoustic searching consumption in this tree is reduced. In this tree, we define

- Root Node (RNode) as the topmost node. It is a virtual node just containing pointers to its child nodes. The RNode has no parent node.
- Phone Node (PNode) as a node that describes a certain ISDP of a syllable. It contains information of the ISDP and pointers to its child nodes.
- Syllable Node (SNode) as the bottommost node, it is the leaf node (LNode) of the tree. It just contains information of the syllable which consisted of the ISDPs from RNode to its parent with the exactly sequence. An LNode has no child node.

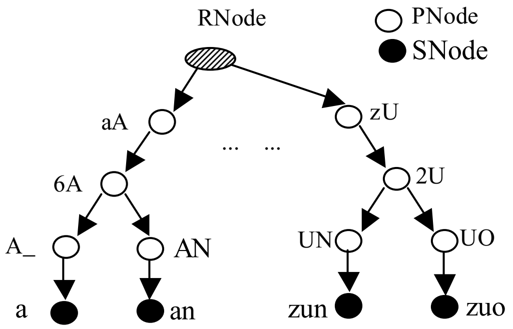

Figure 2. A partial ISDP-Syllable Search Tree

The frame-synchronous network search algorithm is used for the ISDP decoding. Given a specific utterance A, we first set the current position at RNode, As the frame vectors is passing by the algorithm, the ISDP-syllable search tree provide the pruning information, and the path is going from RNode to SNode according to the tree. When a certain SNode is reached, the syllable corresponding to that SNode is got as the recognition result. The decoding process should also obtain the IDSP sequences of each frame vector, we can retrieve this by trace back the search tree from the SNode to the RNode.

5. EXPERIMENTAL RESULTS

A continuous Chinese speech corpus from 863 materials is used in the following experiments. The corpus contains 13 speakers' data and there are 520 utterances available for each speaker. All the recorded materials are obtained in a low noise environment through a close-talk noise-canceling microphone. Ten speakers' data are used as the training database for ISDPMs; the remaining part is used for testing. They are digitized at a sampling frequency of 16KHZ. A 32ms Hamming window is applied to each frame of speech. And then the cepstral coefficients derived from 16-order LPC 16 are extracted every 16ms. Each ISDP is divided into 2 states.

To build the initial ISDPMs we should have the utterance with ISDP level label. But the utterance we used here is only labeled with the syllable level and the work of labeling the utterance to ISDP level by hand is an onerous job. Here we use the syllable label to generate ISDP label as follows: for each syllable i, we count the state number t_i of the component ISDPs (each ISDP is consisted of 2 states), then the NLP algorithm is adopt to divided the syllable into t_i segments, we use the divided segments label as the ISDP label. Certainly this treatment may have low accuracy, but by this way we can establish the initial ISDPMs. We can iteratively train the model several times to obtain good performance.

The follow table gives the accuracy of ISDPMs with different mixture number and iterative times.

Table5. Performance of ISDPMs

Mixed number	**Iterative times**	**Accuracy Rate (%)**		
		Top 1	**Top 5**	**Top 10**
8	Initial	62.54	83.33	87.62
	1	68.27	86.02	90.29
	2	71.68	86.87	91.78
	3	73.39	88.23	93.03
16	Initial	63.71	84.24	88.60
	1	69.38	86.32	91.21
	2	72.17	87.15	92.96
	3	74.28	88.74	93.29

From the table we can see that the iterative training of ISDPMs can reach much higher accuracy than the initial model. The number of mixed gaussian densities for each state if also an important Criterion of the ISDPMs, which can make the error rate be reduced by about 3.3% when increase the mixed number from 8 to 16.

To illustrate the performance of ISDPMs, we train the MGCPMs as the contrast experiment. The SRU of MGCPMs is syllable; each syllable is divided into 6 states. The following table shows the perfromance of MGCPMs.

Table6. The performance of MGCPMs

Condition	**Accuracy Rate (%)**		
	Top 1	**Top 5**	**Top 10**
4 Mixtures, 3 iteration	74.35	89.15	93.78

From the table5 and table6, we can see that the MGCPMs with 4 mixed gaussians for each state can cause only about 2.7‰ reduction of the error rate. The MGCPMs of 4mixtures have nearly 10,000 guassians while the ISDPMs have only about 4,000 gaussians. Thus the complexity of ISDPMs can be far less than the MGCPMs.

6. SUMMARY

In this paper the ISDPMs for Chinese speech recognition is proposed and studied. The experimental results show that the ISDPMs can reduce the computational complexity during the recognition process with only a little degradation in accuracy compared to the MGCPMs. The ISDPMs can be utilized for the applications with low computational power. In our future study, we will extend the ISDPs with more detail contexts and do more experiments on the phonetic model such as di-phone or tri-phone model.

7. REFERENCES

[1] L.R.Rabiner, “A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition”, *Proc. IEEE*, 72(2):257-286, Feburary 1989

[2] 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, P.R.China, Oct 1996

[3] Zheng F., Mou X.-L., Xu M.-X., Wu J.,Song Z.-J., “Studies and Implementation of the Techniques for Chinese Dictation Machines,” J.Software, 10(4):436-444, April 1999

[4] Zheng, F., Song, Z.-X, Xu, M.-X, “EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine”, EuroSpeech’99, Vol.2, pp.819-822, Budapest, Hungary, Sept.1999

[5] Gao S., Xu B., Huang T.Y., “Class-triphone Acoustic Modeling Based on Decision Tree for Mandarin Continuous Speech Recognition”, ISCSLP’98, ASR-A1, Singapore , 1998

[6] Viterbi, A.J., “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. on IT*, 13(2), Apr. 1967

[7] Xu, M.-X,. Zheng, F., Wu, W.-H, A fast and effective state decoding algorithm, EuroSpeech’99, Vol. 1, pp.187-190, Budapest, Hungary, Sept. 1999