Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

APPLYING CONDITIONAL RANDOM FIELDS ON CHINESE SYLLABLE RECOGNITION

JIE LI, XUAN WANG, YI YANG
Intelligent Computation Research Center, Harbin Institute of Technology Shenzhen Gradual School
E-MAIL: freespace_233@yahoo.com.cn, wangxuan@cs.hitsz.edu.cn

***Abstract*—Hidden Markov model (HMM) is successfully used in speech recognition. However, there is an unavoidable flaw in assuming strong independence for sequences labeling in HMM. While Conditional Random Fields (CRFs) can relax this assumption for HMM, and can also solve the label bias problem efficiently. In this paper, we investigate CRFs for Chinese syllable recognition in continuous speech due to its advantages. The experiments show that the syllable label CRF is able to achieve performance comparable to phone-based HMM.**

***Keywords*—Chinese syllable recognition, CRFs, HMM**

I. Introduction

Hidden Markov model (HMM) has been widely used for acoustic modeling in speech recognition. Regardless of its great success, there are some defects to be modified or improved e.g. strong independent assumptions. In fact, the labeling sequences which are called frames in speech recognition are dependent rather than independent when the speech signal is continuous. Due to this characteristic, many researchers in the field of speech have switched from generative models to discriminative models.

Recently, Conditional Random Fields (CRFs) [1] are becoming more and more popular as a sequence labeling model, since they avoid independence assumptions in HMM. Furthermore, it has the ability to incorporate a rich set of non-independent features. In the tasks of part-of-speech tagging, named-entity recognition and shallow parsing, CRFs perform even better than HMM. However, research in the field of CRFs on speech recognition is not sufficient. Since A. Gunawardana et. al. [2] firstly approached CRFs for phone classification in 2005, many researchers focus on applying CRFs in speech recognition. J. Morris et. al. [3] combined phonetic attributes using CRFs and detector-based CRFs for phonetic recognition. In their experiments, they make use of the TIMIT acoustic phonetic corpus for training and testing. While there are totally 61 possible phone labels in TIMIT, the authors used a reduced phoneme labeling for TIMIT of 39 possible outputs instead of the full 61 phone labels [4] and got the best phone accuracy of 70.40% reported in [3].

In this paper, we approach applying CRFs on Chinese syllable recognition in continuous speech. In Chinese, syllable is often a better and easier recognition unit, and there are totally 408 syllables. So in the first step of Chinese speech recognition, the speech is recognized as a group of corresponding syllables. Thereafter these connective syllables are transformed into Chinese sentences using language model. We focus in this paper on the acoustic modeling, and each syllable is a unique label, so there are totally 408 types of labels in the problem. At first, the speech to be recognized is divided into many frames through signal preprocessing. The frames thus obtained can be taken as observation sequences, which are to be labeled by the syllables.

The paper is organized as follows. In the next section, we present an overview of CRFs. Section III describes application of this model in Chinese syllable recognition. The training method and the feature templates of our CRF model are described in Section IV. In section V, the experiments are described in detail. Finally, the conclusions and future work are presented in Section VI.

II. Conditional Random Fields

Suppose that $X=(x_1,x_2,...,x_n)$ is a random variable over data sequences to be labeled and $Y=(y_1,y_2,...,y_n)$ is a random variable over corresponding label sequences, according to CRFs, the random variable X and Y are jointly distributed, and $p(Y\mid X)$ denotes the conditional probability distributions of label sequences given input sequences. Thus, CRFs can be described by an undirected graph $G=(V,E)$ such that $Y=(Y_v)_{v\in V}$, Y is indexed by the vertices of G. Then (X,Y) is a conditional random field , in case when conditioned on X, the random variables Y_v follow the Markov property with the undirected graph: $p(Y_v\mid X,Y_w,w\neq v)=p(Y_v\mid X,Y_w,w\sim v)$, where $w\sim v$ means that w and v are neighbors in G. In a word, a CRF is a undirected graphical model globally conditioned on the observation sequence.

A CRF on (X,Y) is specified by a vector f of local features and a corresponding weight vector λ. Lafferty et al. [1] define that the probability of a particular label sequence Y given observation sequence X is a normalized product of potential functions which is presented in (1):

$$p_\theta(y\mid x)\propto \exp(\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)) \tag{1}$$

Where $t_j(y_{i-1},y_i,x,i)$ is a transition feature function of the entire observation sequence, as well as the labels at positions i and $i-1$ in the label sequence; $s_k(y_i,x,i)$ is a state feature function of the label at position i and the observation sequence; in order to express conveniently, we usually use

$$F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x,i) \tag{2}$$

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

Where $F_j(y,x)$ denotes the feature functions and each $f_j(y_{i-1},y_i,x,i)$ in (2) is either a state function $s_k(y_i,x,i)$ or a transition function $t_j(y_{i-1},y_i,x,i)$. So equation (1) can be written as:

$$p(y\,|\,x,\lambda)=\frac{1}{Z(x)}\exp(\sum_j \lambda_j F_j(y,x)) \qquad (3)$$

Here $Z(x)$ is a normalization factor and $\theta=(\lambda_1,\lambda_2,...,\lambda_n)$ is the parameter vector to be estimated from training data.

CRFs are trained by maximizing the log-likelihood of a given training set $T=\{(x_k,y_k)\}_{k=1}^{N}$ which is equivalent to the equation in (4) derived from of equation (3):

$$\hat{y}=\arg\max_y p(y\,|\,x,\lambda)=\arg\max_y \lambda\cdot F(y,x) \qquad (4)$$

The training of CRFs is very important, and there are many methods for training, including preconditioned conjugate gradient, limited-memory quasi-newton, voted perceptron and etc.. A more detailed training methods of CRFs can be found in [5].

As compare to generative models (HMM-like models), the class of CRFs is more expressive, because it allows arbitrary dependencies on the observation sequence. In addition, the features do not need to completely specify a state or observation, so one might expect that the model can be estimated from less training data. Another attractive property is the convexity of the loss function; CRFs share all of the convexity properties of general maximum entropy models [1].

III. APPLICATION OF CRFS IN CHINESE SYLLABLE RECOGNITION

As doing speech recognition with HMM, we perform signal processing by extracting the acoustic features of each utterance in the training speech corpus first, and the extracted acoustic features in continuous speech can be thought as observation sequence to be labeled. So the labeling information (syllable) of the corresponding utterance can be combined to train CRFs. And finally, we can do testing with the trained models.

Generally, the speech signal should be divided into many frames in order to process conveniently. In our experiments, the frame is set by 24ms, and the frame shift is 12ms. After that, Linear Predictive Coefficients (LPC) or Mel Frequency Cepstral Coefficients(MFCCs) acoustic features are extracted from each frame. We just extract the 12-dimentional MFCCs, the 12-dimentional delta coefficients and the 12-dimentional acceleration coefficients of MFCCs, totally 36-dimentional features of MFCCs. These acoustic features of each frame can be seemed as the basic unit to be recognized, and each frame corresponds to a unique syllable except for the frames that contain no sound information. So each utterance in the speech corpus is composed of many frames that are to be labeled by all the syllables in Chinese. During the training perioid of CRFs, the lable of each corresponding frame in the training speech corpus should be known. In our experiments, all the speech for training and testing is already segmented. Thus, the problem of Chinese syllable recognition is turned to the problem of labeling sequences with the fixed types of labels.

There are a lot of statistcal models (Bayesian Network, SVM, ANN...) for labeling sequences, We apply CRFs in this paper to label sequence with Chinese syllables. The training and testing course in CRFs model is presented in Section IV in detail.

IV. CRFS MODEL IN THE EXPERIMENTS

In this section, we would focus on the training of CRFs in the experiments. In CRFs, choosing appropriate feature sets is an important issue. All kinds of features can be added into the CRFs model very flexibly, and choosing an efficient feature set would be the key of an experiment. In CRFs, we don't need to assume the independence of sequences, and the context-dependent sequences (frames in speech recognition) must be considered. The context stands for a windown which contains some frames in order to consider the relations among them. Theoretically, the larger the window, the more context information can be acquired, but as the window becomes larger, it greatly reduce the efficiency of the training procedure. On the other hand, if the window is too small, the features would not be used sufficiently, and some important information may be lost. Based on the analysis, the better window of sequences should be no more than 2 [5]. Besides the window, a good defined feature function is also important. Based on the characteristic of speech signal, we have defiend the following 10 feature functions.

$$f_s(y_i,x_i)=\begin{cases}x_i & y_i=s\\0 & others\end{cases} \qquad (\text{F1})$$

$$f_s(y_i,x_{i-1})=\begin{cases}x_{i-1} & y_i=s\\0 & others\end{cases} \qquad (\text{F2})$$

$$f_s(y_i,x_{i-2})=\begin{cases}x_{i-2} & y_i=s\\0 & others\end{cases} \qquad (\text{F3})$$

$$f_s(y_i,x_{i+1})=\begin{cases}x_{i+1} & y_i=s\\0 & others\end{cases} \qquad (\text{F4})$$

$$f_s(y_i,x_{i+2})=\begin{cases}x_{i+2} & y_i=s\\0 & others\end{cases} \qquad (\text{F5})$$

$$f_{s,s'}(y_{i-1},y_i)=\begin{cases}1 & y_{i-1}=s \text{ and } y_i=s'\\0 & others\end{cases} \qquad (\text{F6})$$

$$f_s(y_i,x_{i-1},x_i)=\begin{cases}x_{i-1}\cdot x_i & y_i=s\\0 & others\end{cases} \qquad (\text{F7})$$

$$f_{s,s'}(y_{i-1},y_i,x_i)=\begin{cases}x_i & y_{i-1}=s \text{ and } y_i=s'\\0 & others\end{cases} \qquad (\text{F8})$$

$$f_{s,t'}(y_i,x_i,t)=\begin{cases}x_i & y_i=s \text{ and } t=t'\\0 & others\end{cases} \qquad (\text{F9})$$

$$f_{s,t'}(y_i,x_{i-1},x_i,t)=\begin{cases}x_i\cdot x_{i-1} & y_i=s \text{ and } t=t'\\0 & others\end{cases} \qquad (\text{F10})$$

In the above equations, x_i denotes the acoustic 36-dimentional features of the current frame, x_{i-1} , x_{i+1} , x_{i-2} , x_{i+2} respectively denote the acoustic 36-dimentional features of the last frame, the next frame, the second frame before the current frame and the second frame after the current frame. y_i denotes the label (which syllable) of the current frame. y_{i-1} denotes the label of the last frame, t denotes the current time information of the frame. From the above 10 feature functions, we combine some of them to form 4 templates, and perform the experimet B in section V, finally T4 was selected as the template in our CRFs model.

TABLE I. TEMPLATES

Template No.	Combination	Meaning
T1	F1+F2+F4+F6	Transition features Window size is 1
T2	F1+F2+F3+F4+F5+F6	Transition features Window size is 2
T3	F1+F2+F4+F6+F7+F8	Transition features Window size is 1
T4	F6+F9+F10	Transition features Time considered

In most CRFs models, the value of defined feature function is binary, but in speech, the acoustic features are made of 36 or more dimentional features, in each dimention, the value of the feature is real, so the number of observation sequences' type is very large, For the large number of features the technology of vector quantization should be adopted, and selecting the representive and less redundant features is very important. Generally, we just choose features by counting. The counting procedure is applied only to choose the features, and can be realized as constant, which can also be written as $F = \{f \mid \sum_{x,y} f(x,y) \geq K, f \in C\}$, C is the feature space, K is a constant. This method is very simple and practical, although there may be some redundant features. In the experiment, we define the K equaling to 1.

After the feature template is confirmed, the parameter vector $\theta = (\lambda_1, \lambda_2, ..., \lambda_n)$ can be estimated by maximizing the log-likelihood of the training data with the iterative scaling algorithms, such as Newton, BFGS and L-BFGS. Among them, L-BFGS is an efficient method, and we choose it to train models. However, it would still take some time to train the model. As for the testing of CRFs, formula (3) is used to search the most possible labels of sequence with Viterbi algorithm.

V. EXPERIMENTAL SETUP

A. Overview

In this section, three experiments are presented which are new applications of CRFs to do syllable recognition in Chinese continuous speech. For general consideration, we choose six datasets for training and testing. These datasets are all from the mandarin standard corpus of Chinese National 863 Project, which includes almost 85,000 utterances, whose sampling frequency is uniformly 16 kHz and coding bit is 16. They have been recorded by 83 male and 83 female. However, since training CRFs needs the label information of each utterance, we just use 4800 utterances, which were recorded by 11 different persons and have been labeled all by ourselves manually, to do experiments. The detailed situation of the six representative datasets is shown in table II.

TABLE II. THE SITUATION OF CORPUSES

No.	Corpus I	Corpus II	Corpus III	Corpus IV	Corpus V	Corpus VI
F00C	420	95	400	50	100	20
F01A	400	50	400	50	100	20
F02A	160	10	--	--	--	--
F04A	80	5	--	--	--	--
F11C	420	95	400	50	100	20
M11B	422	95	400	50	100	20
M12C	423	95	400	50	100	20
M13C	420	95	--	--	--	--
M14C	422	95	--	--	--	--
M15C	410	70	--	--	--	--
M16C	423	95	--	--	--	--
Total	4000	800	2000	250	500	100

Notice that in all of the experiments below, we use the uniformly common 36-dimentional MFCC acoustic features.

The remainder of this section proceeds as follows. In part B, we first design an experiment of 61-syllable recognition to find a good template for our CRFs. In Sub-section C, recognition extends from 61-syllable to 265-syllable, meanwhile, HMM is approached to do the experiment for comparison. Finally, another experiment of 61-syllable recognition with different training and testing datasets is set up to shown the convergency of CRFs.

B. 61-Syllable Recognition

Although there are 408 syllables in Chinese, firstly, as the space of features increases, the training for CRFs takes more time. Secondly, when the number of recognized syllables becomes larger, the larger training dataset is needed in order to train each syllable sufficiently. Thus, we set up an appropriate experiment for 61-syllable recognition to find a good template for our CRFs, here Corpus I is chosen for training, Corpus II for testing. As each syllable doesn't occur the same times in the corpus, we arrange them in descending order by the frequency of each syllable occurring in the training corpus and then use CRFs to model the front 61 syllables. As discussed in section IV, templates T1, T2, T3 and T4 are defined to compare the experimental results. The below Fig. 1 describes the number of features generated by these four templates and the precisions of recognition are shown in Fig. 2. Although there are much more features in template T4, the precision of T4 is the highest. Finally we choose template T4 in our CRFs models, and the best result of 61-syllable recognition with template T4 is also shown in table III.

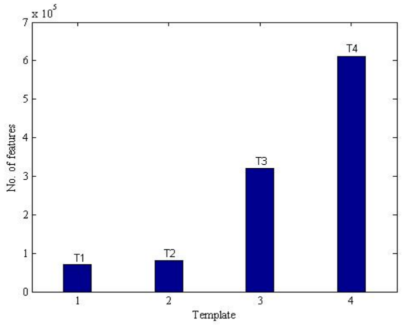

Figure 1. Features generated by different templates

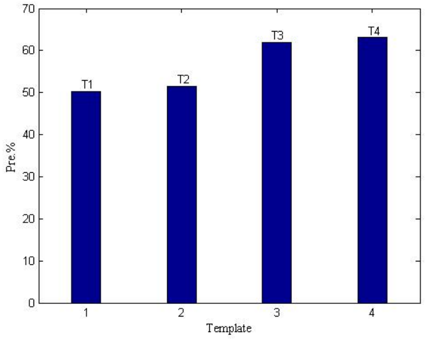

Figure 2. Precisions of different templates

TABLE III. THE RESULTS OF 61-SYLLABLE RECOGNITION WITH T4

Training dataset (Corpus I)	Precision of testing dataset (Corpus II)
F00C	55.15%
F01A	61.11%
F02A	63.83%
F04A	39.29%
F11C	63.92%
M11B	59.46%
M12C	66.57%
M13C	77.94%
M14C	62.19%
M15C	61.52%
M16C	67.88%
Total	64.12%

C. 265-Syllable Recognition

This experiment is set up to show the performance of CRFs in the recognition of large syllables. Similar to part B, the 265 most frequently occurring syllables are modeled by CRFs. Besides this, we also use HMM to do the same task with HTK [6]. The results of these two models are described in table IV. Here Corpus V is used for training and Corpus VI for testing.

TABLE IV. THE RESULTS OF 265-SYLLABLE RECOGNITION

Model	Trained Syllables	Correction
CRFs	265	48.80%
Monophone-HMM	408	48.13%
Triphone-HMM	408	57.15%

According to the results, the model of CRFs can be comparable to monophone-HMM, although the result of triphone-HMM is superior to CRFs. However, in this experiment, we have just trained 265 syllables, and the training of HMM uses a smaller unit, phone (initial and final in Chinese) to recognize, while syllable is just processed as the basic recognition unit in our model of CRFs here.

D. The Convergency of CRFs

Training is always a very important issue in CRFs, and it becomes a problem especially when the feature space is large. Here we approach L-BFGS to train CRFs, and design two models of CRFs to do 61-syllable recognition, in one model, Corpus I for training and Corpus II for testing, in another, Corpus III for training and Corpus IV for testing. Although the training and testing datasets for the two models are different, it would not influence the purpose of this experiment since we just focus on the convergency here. The results are shown in Fig. 3.

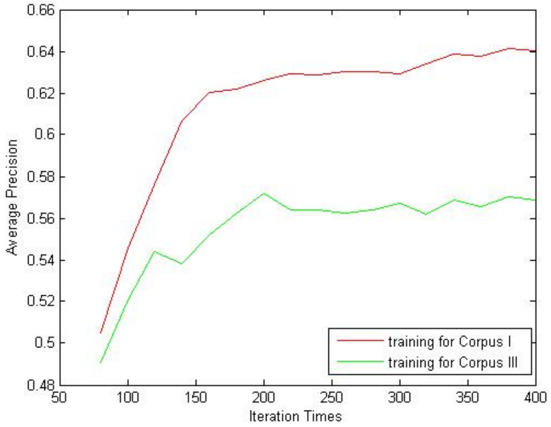

Figure 3. The convergency of two CRFs models

From the Fig. 3, we can see very clearly that for the purpose of convergence, the iteration times for both of the two models should be more than 200. when the number of iterations are less than 200, especially less than 150, it seriously influence the testing result. Generally, since the features of CRFs in speech is a bit large, it would take much time to get a completely convergent model. However, according to this experiment, we can get the experience that a good precision of recognition could be acquired when the iteration times are between 200 and 400.

VI. CONCLUSION AND FUTURE WORK

In this paper, we apply the model of CRFs to do syllable recognition in Chinese continuous speech. Three experiments are designed to show the performance of CRFs in Chinese syllable recognition. At first, an experiment of 61-syllable recognition is set up to find a good feature template in the CRFs model. Based on the confirmed template, we design another experiment for larger syllables' recognition (265-

syllable) and acquire a good result which could be comparable to HMM just using the syllable rather than the phone as the basic recognition unit. Finally, we present the convergency of CRFs through the experiments and discuss the problem of iteration times with CRFs applied to syllable recognition.

However, there are still some important issues to be tackled for the future researches. First, the large feature space in speech recognition with CRFs should be reduced and good feature templates are desirable. Secondly, there is a need of some new methods to be proposed for acquiring effective features, for example, some researchers have proposed the multilayer perceptron ANN to detect effective phonological features [7]. So we can investigate these methods to improve the training speed of CRFs. Second, we can use a smaller recognizing unit (called consonant and vowel in Chinese) to improve the precision of recognition.

VII. Acknowledgment

The authors would like to thank Lu Li and Farooq for useful discussions of the paper. This work is supported by the National High-tech R&D Program of China (863 Program, No. 2007AA01Z194).

References

[1] J. Lafferty, A. McCallum, and F. Pereira, “Conditional—Random fields: Probabilistic models for segmenting and labeling sequence data,” in Proco. ICML, 2001, pp. 282–289.

[2] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, “Hidden conditional random fields for phone classification,” in Proc. Interspeech, 2005.

[3] J. Morris, and E. Fosler-Lussier, “Conditional Random Fields for Integrating Local Discriminative Classifiers,” IEEE Trans. Audio, Speech, and Lang. Process., vol. 16, no. 3, pp. 617–628, March 2008.

[4] K. Lee and H. Hon, “Speaker-independent phone recognition using hidden Markov models,” IEEE Trans. Acoust., Speech, Signal Process., vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[5] F. Sha and F. Pereira, “Shallow parsing with conditional random fields,” in Proc. of HLT, NAACL, 2003.

[6] S. Yong, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book(for HTK Version 3.2). Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2002.

[7] S. King and P. Taylor, “Detection of phonological features in continuous speech using neural networks,” Comput. Speech Lang., pp. 333-353, 2000.