Research on Segment Acoustic Model Based Mandarin LVCSR

Wenju Liu, Yun Tang, and Shouye Peng

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, P.O.Box 2728, China
{lwj,ytang,sypeng}@nlpr.ia.ac.cn

Abstract. SM has shown a better performance than HMM in connected word recognition system; however, no reports we have read show that SM has been applied in LVCSR as decoding acoustic model because of the restriction of its complexity. We have preliminarily built a SM based mandarin LVCSR system which adopts CART and global tying to tie the parameters in the triphone models and the fast SM algorithm, CF algorithm and two-level pruning to enhance the speed of decoding. The system achieves 87.09% syllable accuracy in Test-863 data corpus within 4 real times. We believe SM offers an alternative choice for LVCSR system though further research for its fast algorithms by rational utilization of its structure information.

Keywords: Segment acoustic model, HMM, SSM, LVCSR, CF algorithm, Parameter Tying.

1 Introduction

Segment Model (SM) [1] is a family of methods which model and decode the observation sequence in a way of segment style compared with HMM whose decoding is in a frame-based style. This characteristic can overcome some limitations in HMM, such as the feature vectors are conditional independent given the state sequence; non-stationary observation sequence is modeled by a piecewise state sequence [2]. SM is totally different from HMM on its segmental decoding way and with potential to finish some works effectively that are in nature difficult in HMM based system, such as integrating more segmental information in decoding process, producing the n-best list during decoding process etc. Experiments prove SM has a better performance than HMM in connected word recognition task [3,4]. But we haven't read any reports to verify the performance of SM in LVCSR system because of the restriction of its complexity. This is our primary motivation to make such a system.

This paper is organized as follows. A brief introduction to stochastic segment model (SSM) [3] and the fast SM is given in the next section. Then in section 3 we will introduce the primary characteristics of our SM based system: parameter tying, CF algorithm and two-level pruning. Section 4 shows the experimental results of SM compared with a HMM system. Finally, conclusions and our future works are given in section 5.

W. Yu, H. He, and N. Zhang (Eds.): ISNN 2009, Part II, LNCS 5552, pp. 928–935, 2009.

2 Fast SSM

2.1 Introduction to SSM

The acoustic model in our system is SSM which represents observation sequence by a fixed length region sequence. A resample function is used to map the variable length segment x_1^N to the fixed length frame sequence y_1^L. Usually, an $L \times d$ dimensional multivariate Gaussian distribution is used to model the segment, where L is the fixed length of frames that are so called "region", and d is the dimension of the feature vector in each frame. The log conditional probability of a segment x_1^N given model α is:

$$\ln[p(x_1^N \mid a)] = \sum_{i=1}^{L} \ln[p(y_i \mid a, r_i)] \cdot \tag{1}$$

Use The decoding process for SSM in sentence x_1^T is:

$$J_m^* = \max_{\tau,\alpha}\{J_\tau^* + \ln[p(x_\tau^m \mid \alpha)](m-\tau) + \lambda \ln(P_S(x_\tau^m \mid \alpha)) + \ln[P(\alpha)] + C\} \cdot \tag{2}$$

$$\phi_{\mathrm{m}}^* = \arg\max_{\tau,\alpha} J_{\mathrm{m}}, J_0^* = 0, 1 \le m \le T,$$

where J_m^* is the accumulated score of the best reference model sequence that ends at time point m ; $p(x_\tau^m \mid \alpha)$ is the likelihood score for segment x_τ^m ; $P_S(x_\tau^m \mid \alpha)$ is the segmental level information such as duration distribution; and C is the penalty factor for each segment.

The final solution for this best path is ϕ_T^* and the path can be retrieved from the end point T of the observation sequence.

2.2 Fast SSM

The high complexity of SM is due to the evaluation of segment score which cannot be decomposed and the intermediate information of score evaluation is not shareable between different segments, even for the case in which two segments only differ in one frame. Most of works accelerating SM are focused on efficient pruning algorithms [5]. These algorithms speed up SM greatly, but they are still far slower than HMM, since the computation of these algorithms is based on segment while HMM on frame.

SSM can be put into the constrained mean trajectory segment model (CMTSM) [1]. The computation on a region is independent from other regions given the segment and only relates to the observation vector and the position of region in the CMTSM. The segment score is the summation of region scores in a linear way without complex operations, e.g., dynamic time warping. The key of fast SM is to decompose the computation on segment into the computation on a series standard region models. Those scores of standard region models can be shared between different segments which are only different in a small part of observation vectors. In fast SSM, the parameters of region models are fixed in order to share the region scores in different segments. It means both length of region sequence and region model parameters will

keep unchanged whenever the length of segment is changed, and the variable length segment will map to the region models by a linear resample. The score of the observation vector for a standard region model can be shared by different segments with mapping between this observation vector and this region model. At each time point, we will only compute the scores of the current feature vector for all of the active region models instead of the active segment models. Though the algorithm of fast SSM is segment based, the main computation, the measure of probability distribution, is frame based. The fast algorithm will decrease the computing time cost of SSM to the tenth of the original one's. The run time of SSM in digit string recognition task falls to the same level as HMM by using the fast algorithm, though the former is a monophone based system and the later is a triphone based system [4]. Fast SSM paves the way for applying SSM to LVCSR system in current computation environment.

3 SSM Based LVCSR System

3.1 Parameter Tying

In LVCSR, the number of context-dependent models for mandarin, i.e. triphone context, is very large. Therefore parameter tying techniques are required to cut down the number of parameters, and hence reduce the computation complexity and improve the robustness of model. Our parameter tying is twofold: CART tying and global tying. First, CART will cluster the regions across models which are derived from the some monophone with different context. The regions for tying together are in the same region position of segment model. Right-sized tree algorithm [2] is used to automatically determine the size of the CART. The question set in system is similar to [6], except that we don't use the tonal questions. In the second step, the region models will be merged again in the whole region set. After the first step, the region models are in an optimal balance between modeling ability and complexity in the condition of the same region position. However, the parameter tying of region models in different region positions and different phoneme classes are not considered. For example, the region models neighboring in position may be tied together to show the fact that they represent the stable physical vocal tract. A simple and effective way we adopted in experiments is to use the bottom-up strategy to merge the region models in the whole region set. If the reduction of probability likelihood after two nodes merged is less than the pre-set threshold, these two nodes will combine to form a new node; otherwise, the merging process will stop. The first step is to get a robust model and the second is for getting an efficient model.

3.2 CF Algorithm

In HMM, a state will automatically decide the stopping point by competing with other states. In SSM, a segment will find the best boundary by competing with other segments and the segment will not know its best boundary before the segment score is measured. So it needs to compute the hypothesis segments from the same start time point with different length, from the minimum length to the maximum length, and most of these extensions are useless. In our task, we propose a Coarse to Fine

Extension algorithm (CF) to alleviate such aimless extensions. The CF algorithm has two phrases, coarse extension phrase and fine extension phrase. In LVCSR, The segmentations of the hypothesis candidates near the right segment are similar, which only differs from others by one or two frames at the beginning or the ending of the segment. The CF algorithm is developed from this observation. During decoding, the CF limits the hypothesis segments ending on every S frames from the start point in Coarse extension phrase; before we form the expanding set spread from point i, we will find out the point $\hat{j}$ with the maximum likelihood to i and expand the neighboring $2(S-1)$ frames near $\hat{j}$. This is the Fine extension phrase. Details of algorithm，see the following：

Coarse to Fine extension algorithm:

1. $i = 0$, candidate set in 0 point: $\Omega(0) = \{"sil"\}$, goto 3;
2. $i{+}{+}$, if $i = T$ goto 7;
3. Form candidate set in point i, if $i < S$ goto 6;
4. Fine Phrase: find $\max(j_s, i)$, expanding candidate segments $(j_s \pm d, i)$, $d \in [1, S-1]$;
5. Coarse Phrase: expands candidate segments (i, j_e) from step 3 and step 4, $i < j_e \le EndFrame(i)$ and $(j_e - i)\%S = 0$, goto 2;
6. Expands candidate set from i point to j_e point, $i < j_e \le EndFrame(i)$, goto 2;
7. End.

Here $EndFrame(i) = \min(i + MaxExtensionFrame, T)$.

In our experiments, CF algorithm can effectively reduce the useless extensions and save more than 30% time of computation with minor influence to the results.

3.3 Two-Level Pruning

The decoding algorithm in (2) is a two-level decoding algorithm. Correspondingly, there are two levels of pruning in system. The first level is done in the process of counting the scores of segment models with the same start point and end point [5]. The computation of regions in a segment is divided into several phrases and at each phrase a threshold is dynamically set to prune those segment candidates with low log likelihood score. The segments survived from this phrase will be pushed to the candidate set and expanded in the following time points. This pruning level can effectively get rid of most wrong hypothesis segments in the initial stage of the segment score computation. The second pruning level is done before expanding the new hypothesis segments from the candidate set. At each time point, first top N_I (or N_F) candidates are selected from syllable initial models and syllable final

models respectively. Then top N_W candidates are selected from the remains of hypothesis segments, including both syllable initial and syllable final. These ($N_I + N_F + N_W$) candidates will be expanded at this time point. The pruning process is illustrated in Fig. 1.

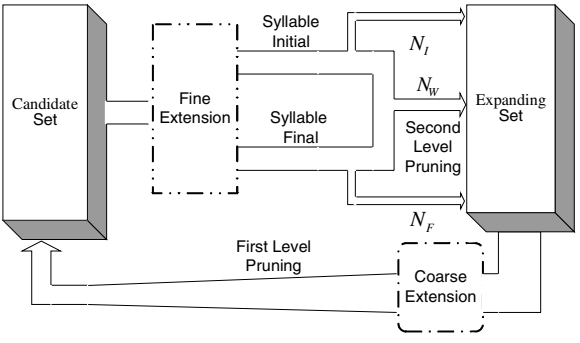

Fig. 1. Two-Level Pruning for SSM System

4 Experiments and Analysis

4.1 Experiment Environment

Phone-set and basic phoneme model: Mandarin is a monosyllabic and tonal language, in which a syllable is composed by syllable initial, syllable final and tone. There are 24 syllable initials and 37 syllable finals in our Mandarin phone-set. Each syllable final has 5 tones. So the number of base phoneme models is 210, including a silence model.

Data Corpus: 83 male speakers' speech data provided by Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development are used for training data and 6 male speakers' for testing data. Both training and testing data are high quality standard mandarin with the accent well controlled. For details see Table 1.

Table 1. Description of speech training and test database

Database	Train-863(55.6 hours)	Test-863(17.1 minutes)
Male Speakers	83	6
Total Utterances	48373	240

Acoustic Feature: 12 dimensions MFCC plus 1 dimension normalized energy and their 1 and 2 order derivative MFCC and energy.

Baseline system: The baseline SSM system is a context-dependent triphone SSM. The pronunciation lexicon is organized in the form of tree structure [7]. The search path begins and ends in silence model. Each segment model has a 15 fixed length region sequence. Each region is modeled by 12 Gaussian mixtures. All SSM systems

in this paper have applied the fast SSM algorithm and two level pruning. Since the complexity of original SSM, we don't measure the running time of original SSM.

To make the experimental comparable, we have developed a CDHMM recognizer as the baseline of HMM by HTK V3.2.1 [8]. The structure of HMM is left to right with 5 states, 3 emitting distributions and no state skipping, except "sp" (short pause) model with 3 states, 1 emitting distribution. Each emitting distribution is modeled by 16 Gaussian mixtures. Both HMM and SM system in this paper share the same training corpus, phone-set and question set in the decision tree.

4.2 Results and Analysis

Our current work is concentrated on SSM acoustic model for LVCSR while the language model is not combined in following experiments. Table 2 gives the baseline results of syllable recognition accuracy in Test-863. In this table, Cor, Del, Ins and Sub are, respectively, the ratio of correct syllable, deletion, insertion and substitution. "Regions" ("States" for HMM) is the number of regions in SSM (or HMM). The unit of time used in results is minute. SMM achieves more 1.5% accuracy than HMM while it also spends more time than HMM.

Table 2. Comparison of HMM and SSM for Test-863

Model	Models	Regions	Cor%	Sub %	Ins %	Del %	Time(min.)
HMM	18364	5068	85.53	14.41	0.06	1.34	15.4
SSM	24180	7983	87.09	12.75	0.16	0.25	94.5

Table 3 gives the results of global tying applied to SSM and the time spent for decoding. β is the threshold of log likelihood reduction for tying two nodes. After global tying, the number of region in model is greatly reduced and the accuracy is downgrade slightly. Experiments show that the global tying works well in this task.

Table 3. Results of Global Tying in SSM

β	Regions	Cor%	Sub%	Ins%	Del %	Time(min.)
800	6866	86.93	12.94	0.13	0.25	88.1
1000	6530	86.74	13.10	0.16	0.22	85.3
1200	6167	86.39	13.42	0.19	0.41	84.6

Table 4 gives the results of syllable recognition accuracy in Test-863 with different step S in CF algorithm. When S sets to 2, the recognition results is slightly changed compared with original one and more than 25% decoding time is saved. When S sets to 3, a little downgrade to the system performance, however, it is not serious and more than 30% time cost is saved.

Table 4. Recognition results with different S by CF algorithm

Regions	S	Cor%	Sub %	Ins%	Del %	Time(min.)
7983	2	87.06	12.78	0.16	0.25	70.8
7983	3	86.77	13.10	0.13	0.22	63.9
6866	2	87.00	12.85	0.16	0.29	69.0

5 Conclusions and Future Works

In this paper, our preliminary work, an SSM based system for LVCSR is proposed. This system can achieve a higher performance than HMM based system. The global tying algorithm in this system can effectively reduce the complexity of SSM with a minor downgrade of the accuracy. A special SM decoding algorithm, CF algorithm, is introduced to alleviate the useless extension during decoding. Coarse phrase offers useful hints for expanding models during fine phrase and fine phrase gives a precise basis for the following coarse phrases. Two-Level pruning can effectively get rid of the impossible hypothesis segments during decoding. Though, the decoding time is still slower than the real time in current system, it makes a firm step towards this goal and the experiments show SSM can be a good alternative acoustic model for LVCSR task. Our future work will focus on refining the CF extension algorithm to enhance the decoding speed and exploring useful segmental information in LVCSR to improve the accuracy of the system.

Acknowledgements. This work was supported in part by the China National Nature Science Foundation (No. 60675026, No. 60121302, No. 90820011), the 863 China National High Technology Development Projects (No.20060101Z4073, No. 2006AA01Z194) and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

References

1. Ostendorf, M., Digalakis, V., Kimball, O.: From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. IEEE Trans. on Speech and Audio Processing 4(5), 360–378 (1996)
2. Huang, X.D., Acero, A., Hon, H.W.: Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, Englewood Cliffs (2001)
3. Ostendorf, M., Roukos, S.: A Stochastic Segment Model for Phoneme Based Continuous Speech Recognition. IEEE Trans. on Acoustic, Speech and Signal Processing. 37(12), 1857–1869 (1989)
4. Tang, Y., Liu, W.J., Zhang, Y.Y., Xu, B.: A Framework for Fast Segment Model by Avoidance of Redundant Computation on Segment. In: ISCSLP, Hong Kong, pp. 117–120 (2004)

5. Digalakis, V., Ostendorf, M., Rohlicek, J.: Fast Algorithms for Phone Classification and Recognition Using Segment-based Models. IEEE Trans. on Signal Processing 40(12), 2885–2896 (1992)
6. Gao, S., et al.: Acoustic Modeling for Chinese Speech Recognition: A Comparative Study of Mandarin and Cantonese. In: ICASSP, Istanbul, pp. 967–970 (2000)
7. Ney, H., Ortmanns, S.: Progress in Dynamic Programming Search for LVCSR. Proceedings of the IEEE 88(8), 1224–1240 (2000)
8. Young, S., et al.: The HTK Book, Cambridge (2002)