IEEE ICASSP, pp.III-1583~1586, June 5-9, 2000, Istanbul, Turkey

STATISTICAL KNOWLEDGE BASED FRAME SYNCHRONOUS SEARCH STRATEGIES IN CONTINUOUS SPEECH RECOGNITION

Zhanjiang SONG, Fang ZHENG, Wenhu WU

Speech Laboratory, Department of Computer Science & Technology,
Tsinghua University, Beijing 100084, P.R.China

szj@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper, we propose a novel and efficient search algorithm for the Continuous Speech Recognition (CSR). The proposed algorithm is on the basis of the traditional Frame Synchronous Search (FSS) algorithm. It makes full use of some statistical knowledge, such as the Differential State Dwelling Distribution (DSDD), as one of the control factors for the state transition. It also incorporates some other rule-based knowledge, such as the pruning criterion based on the dynamic forward prediction and the lexical Word Search Tree (WST), as the search constraint. Experimental result shows that the statistical knowledge based search strategies can improve the performance of the CSR system significantly with an increase of the accuracy by 36.6% compared with the baseline FSS. Also, *EasyTalk*, the Chinese CSR system based on it, has achieved higher recognition accuracy and decoding efficiency.

1. INTRODUCTION

The search process is very important in a Continuous Speech Recognition (CSR) task. For HMMs [6], the Frame Synchronous Search (FSS) [3] algorithm as well as the Viterbi decoding algorithm [8] is commonly used for the time alignment.

The traditional FSS algorithm is originally designed for the connected-word speech recognition, so we make some modifications, such as keeping the backtracking information and the traversed lexical nodes in the partial paths [12], to facilitate its use in the continuous speech recognition. This modified FSS algorithm forms the baseline search algorithm in this paper.

In the implementation of the baseline FSS algorithm, the search is proceeding along the sequence of the speech feature vectors frame by frame. For each partial path stopping at time *t*, the frame at time *t+1* is assumed to belong to any possible subsequent state of it. At time *t+1*, each partial path at time *t* is expanded by appending to it the corresponding frame feature vector as each possible state respectively and thus expanded into several new partial paths with different tailing states. After applying some pruning and merging rules to the paths at time *t+1*, the search process is forwarded by one frame.

The baseline FSS algorithm is concise and effective. While for the recognition tasks of larger vocabulary, the number of the expended search paths grows even more rapidly with time *t* increases, hence a certain threshold has to be chosen to prune some paths of lower scores from time to time. However, stricter threshold may cause some potentially right paths to be discarded beyond recovery, while looser threshold may bring higher burdens to the storage or the search.

In our large-vocabulary, speaker-independent, continuous Chinese speech recognition system *EasyTalk* [12], we introduce a novel and efficient Statistical Knowledge Based Frame Synchronous Search (SKB-FSS) algorithm. On the basis of the fundamental FSS algorithm, SKB-FSS makes full use of some valuable statistical knowledge and some rule-based high-level knowledge for the state transition control and the search constraint, which bring prominent performance to the whole Chinese dictating system.

This paper is organized as follows. In Section 2, we describe the underlying acoustic model that the search strategies are based on. In Section 3, we introduce the SKB-FSS strategies and the search constraint with the Word Search Tree (WST). The experimental results and conclusions are given in Sections 4 and 5.

2. ACOUSTIC MODELING

2.1 Segmental Probability Models (SPM)

Researches and experiments on the distance measure between HMMs have shown that the transition probability matrix plays a far less important role in HMM than the observation probability matrix does [2][4][6]. So the SPM is proposed on the basis of the desertion of the HMM's transition probability matrix. Two instances of the SPM are Center-Distance Continuous Probability Model (CDCPM) and Mixed-Gaussian Continuous Probability Model (MGCPM) [11][13]. These continuous SPMs adopt a left-to-right and non-state-skipping topology. The intra-state feature space is described by Center-Distance Normal (CDN) densities [7] or Mixed Gaussian Densities (MGDs) [11]. The transitions of states are controlled by the high robust Non-Linear Partition (NLP) [1] algorithm which is based on the Equal Feature Variance Sum (EFVS) criterion in the training stage while by the EFVS based search algorithm [9] or the modified Viterbi decoding algorithm [13] in the recognition stage.

The Chinese Dictation Machine Engine (CDME) of the latest *EasyTalk* adopts the 6-state 16-MGD based MGCPM. 419 toneless Chinese syllables are chosen as the Speech Recognition Units (SRUs), the MGCPM achieves a satisfying isolated-syllable based top-10 recognition accuracy of 99.05% for a 30-male training set and the accuracy of 95.65% for a 8-male testing set across the *863-Database* of China [12]. This acoustic model is nearly as good as the traditional HMM in the recognition accuracy while it is much faster and much smaller than the HMM.

2.2 Preparing Recognition Units

Chinese is syllabic and the continuous Chinese speech is always intermittent at the boundaries of words (mostly disyllable words or monosyllable words). Apparently, the intermission infor-mation can be utilized to lighten the load of the search process. Consequently, in order to acquire enough searching efficiency, two steps are adopted in our CDME. Firstly, trying to find as many syllable-separation-points as possible, and secondly, performing efficient state decoding in each segment between the adjacent separating points.

As to the efficient state decoding procedure, we use the SKB-FSS algorithm to be introduced in Section 3. In order to reliably detect all valid speech and detect as many syllable boundaries in valid speech as possible with least errors, we propose a Merging-Based Syllable Detection Automaton (MBSDA) [10] algorithm.

The MBSDA utilizes some time-domain features, such as the short-term frame energy, zero-crossing rate and pitch information, to combine one or several adjacent highly similar raw speech frames into Merged Similar Segments (MSSs). The speech frames in each MSS are regarded to belong to the same phonetic part of a syllable. A Syllable Detection Automaton (SDA) is designed to contain five finite states: silence/noise, Chinese initial, pseudo-silence (exists between the initial and the final of some special Chinese syllables), Chinese final, tail of Chinese final. Figure 1 shows the states and their transitions in the SDA.

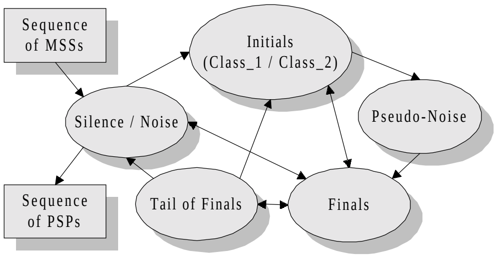

Figure 1. Finite State Transitions in SDA

After processing the consecutive MSSs, the SDA will output a sequence of syllable Putative Separating Points (PSPs), which are either True Separation Points (TSPs) or False Separation Points (FSPs). It is also possible to adjust the SDA parameters so that all the outputted PSPs are TSPs. Each raw speech segment between the two adjacent TSPs is called a Definite Segment (DS), and then the SKB-FSS will be applied to each DS. The possible range of the number of the Chinese syllables contained in a certain DS can also be estimated by counting the number of the connected-final-groups in the DS. The information of the Range-of-Syllable-Number (RSN) acts as another valuable knowledge to perform the path pruning in the SKB-FSS procedure.

3. STRATEGIES FOR SKB-FSS

3.1 Practical Statistical Knowledge in Search

As mentioned above, applying statistical knowledge to the search procedure can greatly improve its performance. Basically, there are two kinds of statistical knowledge to be used in the FSS.

One kind is the probabilistic descriptions based on the pure statistical theory. Typically, for example, we can model the state duration with some probabilistic distributions. In this case, the conditional probability of the duration related to current state is subtracted, as the penalty score, from the total path scores [3][5].

Another kind is the specific rules based on some practically statistical knowledge. For example, according to the statistical histogram of the duration distributions that each state dwells, the states in a search path are allowed to transit or dwell only when their dwelling duration falls into a specific dynamic range in the searching stage. That is, the probabilistic distribution of the state dwelling duration is uniform according to the above case.

The SKB-FSS is mainly based on the specific statistical rues. However, the statistical knowledge adopted in SKB-FSS in not confined in the duration distribution that each states dwells only.

3.2 State Dwelling Distribution (SDD) Based FSS

The SDD is a kind of information that is widely used for path pruning in some kinds of search algorithms. Figure 2 shows the statistical results of the number of dwelling frames in each intra-syllable state. The boundaries of the states are determined by the NLP criterion across the full set of the hand-labeled Mandarin speech database (*863-Database*) of China where the frame size is 32ms and the frame shift 16ms.

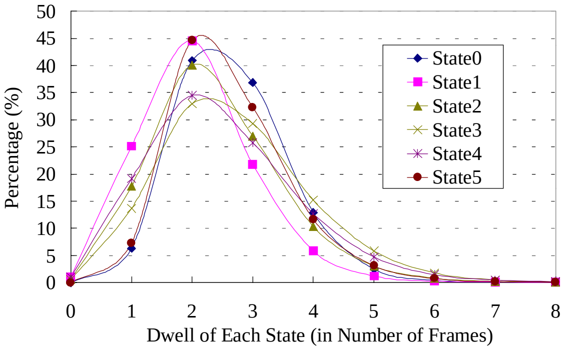

Figure 2. Statistical results on the number of dwell frames in each intra-syllable state (SDD)

According to Figure 2, we come to the conclusion that the most possible duration is 2-frame (32ms) long with the largest proportion, covering 39.6% of the full dwelling set. The duration of 0~5 frames (0~80ms) covers 98.7% while the duration of 1~4 frames (16~64ms) covers 95.0%.

For the sake of simplification and high efficiency in the baseline SKB-FSS, we use an advisable frame number range [$D_{\min}$, $D_{\max}$] instead of the probabilistic distribution of the state duration, to control the state transitions in the search procedure. That is, for the duration *d* of the last state *s* in any partial path in question, if d < $D_{\min}$, the state *s* must dwell, and if d > $D_{\max}$, the state *s* must transit to next state at once, otherwise, the state *s* can either dwell or transit. Obviously, making use of the SDD only can not guarantee the robustness and the performance of the search procedure, because the fixed range may cause the mismatch of the decoded state sequence while the speaking speed varies.

3.3 Differential SDD (DSDD) Based FSS

In order to deal with the variation of the speaking speed and improve the robustness of the SDD in SKB-FSS, we take the variable DSDD-based frame-range instead of the absolute SDD-based frame-range for the state transition control in the SKB-FSS. Figure 3 shows the statistical results of the differential dwelling frame numbers between adjacent intra-syllable states.

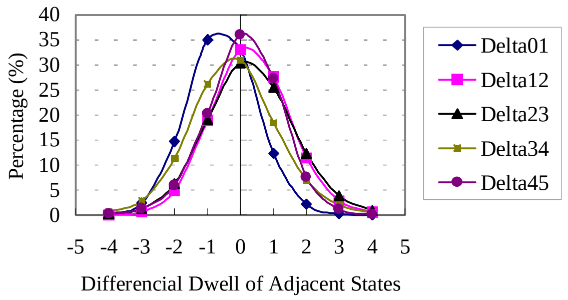

Figure 3. Statistical results of the differential dwelling frame numbers between adjacent states (DSDD)

Figure 3 shows that the most possible differential dwelling frame number between adjacent states is *0*, covering 32.7% of the full differential dwelling set. We can also conclude that the differential dwellings from -4 to 4 frames (-64ms to 64ms) covers 99.8%, while the differential dwellings from -2 to 2 frames (-32ms to 32ms) covers more than 95.4%. That is to say, the duration distributions in adjacent states of the same SRU are not very different from each other.

Consequently, in SKB-FSS, we take $[d_{\min}^{(s)}, d_{\max}^{(s)}]$ as the allow-able differential frame dwelling range for State *s*. Specially, state *0* in each SRU is assigned with a bigger dwelling range to give the initial flexibility for subsequent state transitions. For any other States *s*, the valid number of dwelling frames in State *s* is defined as

$$[D^{(s)} + d_{\min}^{(s)}, D^{(s)} + d_{\max}^{(s)}],$$

where $D^{(s)}$ is the Representative Value (RV) of the last state *s* in current partial path. The RV of state *s* can be defined as the number of dwelling frames in state *s-1*, or defined as the average number of dwelling frames in the traversed states [0, s-1] of the partial path. Comparison results of the SDD and the DSDDs of different kinds of RVs are given in Section 4.

3.4 Pruning Rules in SKB-FSS

Since State *0* in each SRU is given a wider frame dwelling range, the SKB-FSS, where the DSDD is used to control the state transition, can achieve higher robustness in the case that the speaking speed varies much. On the other hand, since the dwelling range of each state *s* is related to its RV representing the search history of current partial path, we can discard some senseless partial paths of invalid dwelling frames as early as possible to improve the search efficiency.

Just as what we have discussed in Section 1, in the process of the baseline FSS, there exists an intrinsic problem that the number of the partial paths exponentially increases with when the frame index (time) increases. Therefore, discarding some very low potential partial paths at any moment will result in a large reduction of the time and space complexities of the SKB-FSS.

Hence, we adopt two kinds of pruning strategies in SKB-FSS to improve its overall performance as follows.

Firstly, whenever current frame is incorporated into all previous partial paths, those newly expanded partial paths with the same tailing states are merged, and only some of those having the highest accumulated scores are kept for future processing in next frame. By discarding some very low potential paths, this kind of pruning rule can largely reduce the search loads and reduce the unwanted disturbance caused by redundant partial paths.

Secondly, the dynamic forward prediction is carried through to judge the validity of each partial path in the search process. According to the TSP and the RSN information provided by the MBSDA mentioned in Section 2.2, we assume that, when one path in the DS reaches the end, its tailing frame must overlap with the final state of a certain syllable. Otherwise, this path is invalid and should be pruned ahead of time.

Figure 4 shows the principle of the dynamic forward prediction based pruning rule for the SKB-FSS.

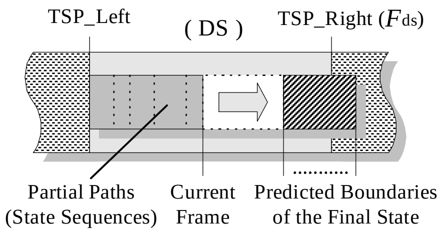

Figure 4. Principle of dynamic forward prediction pruning

From Figure 4 we can conclude that a partial path is reasonable at least before current frame, only if F_{ds}, the right boundary of the DS, falls into the predicted dynamic range $[F_{\min}, F_{\max}]$ that the final state of its tailing syllable can reach according to its RSN information. Otherwise if $F_{\min}>F_{\text{ds}}$ or $F_{\max}<F_{\text{ds}}$, the partial path is invalid and should be pruned at once.

3.5 Search Constraint in SKB-FSS

The lexical Word Search Tree (WST) [12] provides another kind of pruning method. Unlike the above two pruning methods stated in Section 3.4, the WST is related to the vocabulary of the CDME in *EasyTalk*, thus the WST is located between the acoustic layer and the language layer of CDME, or, the WST belongs to the sub-language layer.

The WST is designed to reflect the relations among all the in-vocabulary words so that the redundancy for both the vocabulary storage and the acoustic searching consumption can be largely reduced. In this tree, all nodes except the virtual root node and the leaf nodes are called Syllable Nodes (SN), because each of them represents the shared information of a common syllable out of several words in the vocabulary. The recursive generation rule of the WST can be found in [12].

In the SKB-FSS, when a certain partial path reaches the final state of a certain syllable and is about to transit to another syllable, only the child syllables of the SN corresponding to that tailing syllable in the WST are necessary to be expanded into the path. There is absolutely no need to expand all 418 Chinese syllables into it. This kind of search constraint will improve the efficiency of the search because the storage of the search paths is thus greatly reduced. There is also no loss to the accuracy of the search, since the out-of-vocabulary words have no sense to the ultimate language model of the CSR procedure.

4. EXPERIMENTAL RESULTS

Here we give some experimental results on the comparison of the strategies of SKB-FSS. The experiment is based on the *EasyTalk*, our continuous Chinese speech recognition system. In order to test the validity of above strategies, we disable the use of the language model in *EasyTalk*, and chose a set of utterances with 208 Chinese phrases, each of them consists of 2 to 4 syllables. The SDD and two kinds of DSDDs are tested. These two kinds of DSDDs are of different kinds of RVs. One of them is the DSDD-LST, whose RV is defined as the number of frames in the previous state, while another is the DSDD-AVG, whose RV is defined as the average number of frames in all the previous states.

Table 1 shows the comparison results. In the table, the *n*'th column gives the average syllable recognition accuracy of top-*n* candidates for each syllable in the phrases.

Table 1. Experimental results of different state-transition control strategies

Top-N (%)	1	2	3	4	5
SDD	63.2	67.1	67.6	69.1	69.6
DSDD-LST	77.9	82.8	84.8	86.8	88.2
DSDD-AVG	86.3	89.7	92.6	94.1	96.1

Table 1 shows that the recognition accuracy of the DSDD-AVG increases relatively by 36.6% compared with that of the SDD. In SKB-FSS, there is also a great improvement in the searching speed compared with that of the baseline FSS.

In addition, in current version of the *EasyTalk* with the SKB-FSS (taking the DSDD-AVG as the RV) algorithm and with the Tri-gram based language model enabled, 200 sentences out of the *863-Database* are randomly selected for testing. The Chinese character recognition accuracy reaches 87.6%.

5. CONCLUSIONS

In this paper, we describe the statistical knowledge based frame synchronous search (SKB-FSS) strategies used in the recognition engine of *EasyTalk*. According to the experimental results, we conclude that the appropriate use of some statistical knowledge (such as the DSDD) in the FSS along with some rule-based constraint (such as the dynamic forward prediction and the restriction with the lexical WST) can achieve better overall performance than the baseline FSS does.

On the other hand, the DSDD based state transition rule and the WST based search constraint are still potential to be improved. As to the acoustical modeling, more detailed distribution descriptions for HMM states other than MGCPM deserve further experiments. Furthermore, the use of other kinds of statistical knowledge based RVs in DSDD, such as assigning different dwelling ranges to the states in Chinese initials and finals, is also potential to improve the overall performance of the CSR system.

6. REFERENCES

[1] L. Jiang, "The Study on the Methods and Systems of Speaker-Independent Speech Recognition Based on the Statistical Probability Models". *Master Thesis: Tsinghua University, China*, June 1989 (in Chinese)

[2] B.-H. Juang and L. R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models", *AT&T Technical Journal*, 64(2): 391-408, Feb. 1985

[3] C.-H. Lee and L. R. Rabiner, "A Frame Synchronous Net-work Search Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, 37(11): 1649-1658, Nov. 1989

[4] K.-F. Lee, "Automatic Speech Recognition – The Develop-ment of the SPHINX System", *Kluwer Academic Publishers*, Boston, 1989

[5] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of IEEE*, 77(2): 257-285, Feb. 1989

[6] L. R. Rabiner and B.-H. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, 3(1): 4-16, Jan. 1986

[7] Z.-J. Song, F. Zheng, M.-X. Xu, and W.-H. Wu, "An Effective Scoring Method for Speaking Skill Evaluation System", *Proceedings of EuroSpeech'99*, 1: 187-190, Budapest, Hungary, Sept. 1999

[8] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. on IT*, 13(2), Apr. 1967

[9] M.-X. Xu, F. Zheng, and W.-H. Wu, "A Fast and Effective State Decoding Algorithm", *Proceedings of EuroSpeech'99*, 3: 1255-1258, Budapest, Hungary, Sept. 1999

[10] J.-Y. Zhang, F. Zheng, S. Du, Z.-J. Song, and M.-X. Xu, "The Merging-Based Syllable Detection Automaton in Continuous Chinese Speech Recognition", *J. of Software*, Vol. 10, No. 11, Nov. 1999 (in Chinese)

[11] F. Zheng, X.-L. Mou, W.-H. Wu, and D.-T. Fang, "On the Embedded Multiple-Model Scoring Scheme for Speech Recognition", *International Symposium on Chinese Spoken Language Processing (ISCSLP'98)*, ASR-A2: 49-53, Singapore, Dec. 7-9, 1998

[12] F. Zheng, Z.-J. Song, M.-X. Xu, J. Wu, Y.-F. Huang, W.-H. Wu, and C. Bi, "EasyTalk: A Large-Vocabulary Speaker-Independent Chinese Dictation Machine", *Proceedings of EuroSpeech'99*, 2: 819-822, Budapest, Hungary, Sept. 1999

[13] F. Zheng, W.-H. Wu, and D.-T. Fang, "Center-Distance Continuous Probability Models and the Distance Measure", *J. of Computer Sci. of Tech.*, 13(5): 426-437, Sept. 1998