International Symposium on Communications
and Information Technologies 2004 (ISCIT 2004)
Sapporo, Japan, October 26- 29, 2004

HMM PARAMETER OPTIMIZATION USING TABU SEARCH

Nattanun Thatphithakkul, Supphanat Kanokphara
Speech Technology Section, Information Research and Development Division,
National Electronics and Computer Technology Center
112 Phahon Yothin Rd., Klong 1, Klong Luang, Pathumthani 12120 Thailand
Tel: +66-2-564-6900 Ext. 2257, Fax: +66-2-564-6873
nattanun_t@notes.nectec.or.th, supphanat_k@notes.nectec.or.th

***Abstract*— Hidden Markov Model (HMM) is regularly trained via mathematic functions optimized by gradient-based methods such as Baum-Welch (BW) algorithm. However, optimization from gradient-based methods usually yields only a local optimum. In this paper, Tabu search (TS), an artificial intelligent (AI) technique able to step back from a local optimum and search for other optima, is employed to attack this difficulty. This paper aims to utilize HMM with TS for speaker-independent (SI) continuous speech recognition. The experiment starts from single speaker experiment in order to design and adjust the algorithm. Then, multi-Gaussian context-dependent (CD) model is applied for SI system. The results show the merit of this new algorithm comparing with the original BW.**

I. INTRODUCTION

For almost a decade, HMM takes part deeply in speech recognition field. This is because HMM can normalize speech signal time-variation and characterize the speech signal statistically. In HMM training process, BW method [1] is typically used to optimize acoustic model (AM). By this way, the best AM generating a particular training set can be achieved.

Many researchers agree that optimization from BW algorithm can gain only a local optimum. To reach or get close to the global one, AM is classically initialized by man-labeled transcriptions. Nevertheless, man-labeled transcriptions require many resources (both time and budget) to complete. For this reason, an alternative training method is still essential.

Nowadays, local optimum problem can be solved by many AI techniques such as TS, Genetic Algorithm (GA), Evaluation Programming (EP), etc [2]. As a result, reworking these AI techniques with speech recognition [3] is very interested. TS [4, 5] is chosen in this paper as (1) TS can flee from local optima and keep on searching for a better optimum; (2) TS is simple to implement and understand; (3) TS is flexible and thus can be adapted with BW training easily.

This paper is organized as follows. The next section reviews TS basic knowledge. Then, the hybrid method, HMM with TS, is explained in Section 3. Section 4 illustrates experimental procedures and results while Section 5 concludes the experiment.

II. .FUNDAMENTAL OF TABU SEARCH

As mentioned above, TS is commonly developed for solving local optimization problem. The algorithm keeps historical local optima for leading to the near-global optimum fast and efficiently. The local optima are kept in Tabu List (TL) for making sure that there will be no same local optimum happening again in the process. Another powerful tool in TS is called backtracking. Backtracking process starts from stepping back to some local optimum in TL and then searching a new optimum in different directions. Backtracking is performed when the backtracking criterion (BC) is encountered.

Before TS procedure explanation, the following Tabu components must be defined.

Solution, Search Space, Move and Neighborhood: A solution is an output from a process in the algorithm. Search space is a domain containing all possible solutions. A move is a process creating a new solution from the current solution within search space. Neighborhood is a set of all possible moves from the current solution.

Cost and Objective Function: Cost is a value for judging what solution is better than the others. Objective function returns the solution cost.

Tabu Criterion (TC): To prevent cycling search, some moves should be forbidden under a condition known as TC. Normally, TC will ban local optimal solutions, which are recorded in TL.

BC: In opposition to TC, this condition allows a solution in TC to be a new solution. This usually happens when moving under TC gets stuck in a local optimum.

TS procedure is as follows.

1. Randomize initial solutions and calculate each solution cost from the objective function. Select the highest cost solution as S_0. Then set best_global = S_0. Figure 1 clarifies this step
2. Move randomly around S_0. This neighborhood is defined as S[i] where i = 0, 1, 2, …, N. Calculate each move cost. Select the best move that is not under TC and set it as S_n. Figure 2 points up this step.
3. Keep all S[i] in TL.
4. If S_n cost is better than S_0, $S_0 = S_n$. If S_0 cost is better than best_global, best_global = S_0.

5. Check for BC. If criterion is met, a solution in TL is chosen as S_0. As shown in figure 3, best_cost_neighbor#3 in figure 2 is discarded and S_n moves again from new S_0.

6. Repeat 2-5 until termination criterion is found.

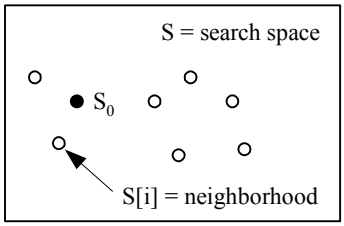

Fig.1. S_0 in search space

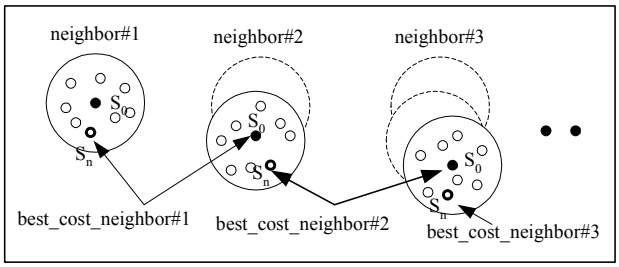

Fig. 2. Tabu procedures.

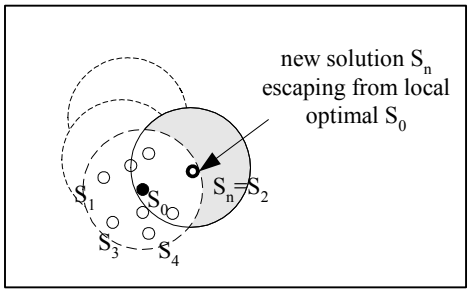

Fig.3. Backtracking process.

III. HMM WITH TS

Generally, BW and TS are similarly used to optimize parameters. BW has strong points that TS does not have and vice versa. BW can optimize parameters with constraints. For each iteration, BW guarantees to give better solution while TS does not. Hence, BW requires less calculation and moves to the final solution faster than TS. Finally, BW can segment speech utterances into quasi-stationary units while TS cannot. The only one strong point of TS over BW is TS can get away from local optima while BW cannot.

Thinking about advantage and disadvantage of both algorithms, combination of these should reasonably produces a better result. BW expedites TS process, segments data and constrains solutions. TS algorithm moves BW solution from a local optimum. To do this, set BW as TS objective function. In another word, BW optimizes the initial solutions from TS. This is similar to old-fashion HMM training that the initial parameters are randomized until the best solution is found. The difference is that TS randomness is not completely blind as TC can prohibit cycling moves and BC can take a solution out of a local optimum.

By introducing TS for BW, the training time may increase. To solve the problem, iteration number and neighborhood member should be kept as low as possible. To decrease iteration number, only some parameters, not all, are randomized. This makes the system converge faster. For neighborhood, the neighbor member is reduced to one. Although the process is simplified as mentioned above, random initialization may take longer time to accomplish the final model. Nevertheless, this longer time is acceptable as the system gives better recognition result without increasing the model complexity and recognition time.

HMM with TS components are redefined as follows.

Solution and Search Space: Solution is acoustic model parameters (means, variances, mixture weights, transitions, etc). Search Space depends on the parameter type. For example, transition must be a real number between 0 and 1. Mean's upper and lower limits are 20 and –20, respectively. Variance is [0.0001, 0.001]. Mean's and Variance's lower and upper limits are decided from the experiment, which is not written here.

Move: For sooner convergence, not all parameters in a solution move. Only some parameters are randomized. These can be means, variances, mixture weights, or their combination, which will be found out in the next section from the experiment. The parameters move continuously with radius

$$rad = \alpha r(X_u - X_l) \qquad (1)$$

where α is a [-1,1] random number, r is a random factor, X_u and X_l are upper and lower boundaries in search space, respectively. Random factor controls perturbation level. For high α, fine search is obtained, albeit time consumption. Low α causes rapid search. From the experiment, r should be 0.001. Then,

$$X_{new} = X_{old} + rad \qquad (2)$$

where X_{new} and X_{old} are new and old solutions within search space, respectively.

Objective Function and Cost: Objective function is BW function where cost is the log probability indicating the model likelihood.

TC: Unlike normal TS, there is no TL here. New solution is acceptable if its cost is higher than the old one.

BC: If there is no solution better than the current one or the cost change is less than 0.01 more than 3 iterations, the backtracking criterion is met.

HMM with TS procedure is as follows.

1. Set model parameters and log probability obtained from BW training as the initial solution and cost, respectively. Then set best_global = S_0.

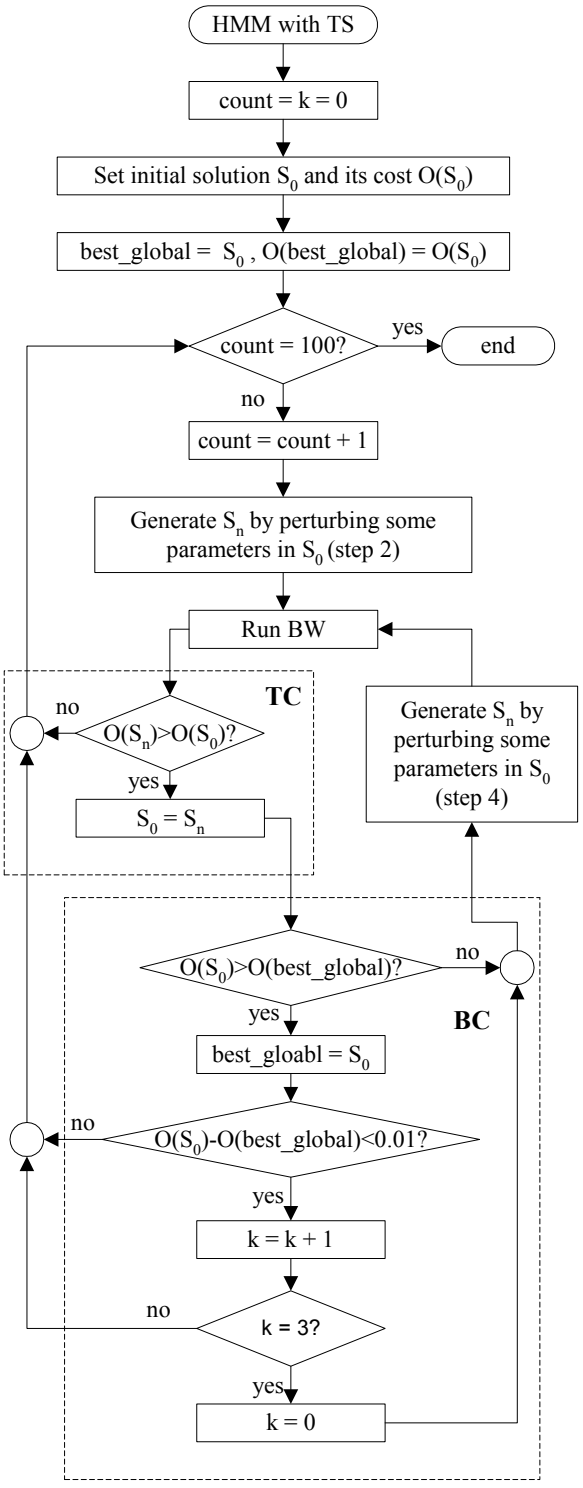

Fig.4. HMM with TS.

2. As BW training needs a lot of time to calculate, only some parameters from S_0 is randomly moved. Set new solution as S_n and calculate its cost. If S_n cost is worse than S_0, do this step again.
3. If S_n cost is better than S_0 then $S_0 = S_n$. If S_0 cost is better than best_global then best_global = S_0.
4. If S_0 is worse than best_global or cost difference between S_0 and best_global is less than 0.01 more than 3 times, randomize some parameters, set it as S_0 and go to step 3. Note that the parameter set randomized in step 2 can be different from here.
5. Repeat 2-4 until number of loops is 100.

Figure 4 is the HMM with TS flowchart. There are two perturbation steps (step 2 and 4). Random parameters are not necessary to be the same in both steps. The system is designed to have 2 perturbation steps because that perturbation in different direction should escape from a local optimum easier is reasonable. Dash-line boxes indicate TC and BC, respectively.

IV. EXPERIMENTAL PROCEDURES AND RESULTS

HTK Toolkit [6] is experimented with 390 Thai phonetic balance (PB) sentences. This PB sentences contain 1478 words, 78 phonemes including short pause and silence unit. The average number of words per sentence is 10. The average number of phones per word is 3.6. All sentences are read by 42 (21 males and 21 females) speakers. The read speech utterances (16 kHz sampling frequency with 16-bit quantization) are parameterized into 12 dimensional MFCC, energy, and their delta and acceleration (39 length front-end parameters). The AM topology is 5-state left-right model. The language model (LM) is trained using back-off bi-gram.

A. One Speaker Experiment

To spot good perturbed parameters for HMM with TS, 390 sentences, one female speaker is selected to train and test with context-independent single Gaussian system. By testing on small close-set experiment, parameters can be tested faster without environment influence such as speaker variation, etc. There are 6 experiment types in this subsection.

1. All transitions.
2. All Means.
3. All variances.
4. Transitions/means.
5. Transitions/variances.
6. No randomization.

"All A" means "A" is perturbed in both steps. "A/B" means "A" is perturbed in step 2 and "B" is perturbed in step 4. No randomization is the baseline system.

HMM with TS are tested with and without LM. Testing without LM explains the real AM result. Testing with LM reveals the improvement limit that LM cannot cover while HMM with TS can. The experiment specifies that the best parameters to be perturbed are transitions at step 2 and means at step 4. For no LM testing, "transitions/means" demonstrates great improvement, 4.65%. Even though LM can support AM error, "transitions/means" still gives 1.25% better than the original one. All results are written in Table 1.

To make sure that HMM with TS is better than original BW training, 200 iterations of HMM with TS and baseline training are observed. Figure 5 illustrates the comparison between best_global's log probability of normal BW and HMM with TS. The original BW training is saturated at the 40th iteration while HMM with TS still can move from the local optimum at 60-80th iterations to better optima.

TABLE I
ONE-SPEAKER EXPERIMENT RESULTS

Experiment types	% Accuracy with LM	% Accuracy without LM
All transitions	97.78	57.73
All means	97.35	53.56
All variances	97.78	57.73
Transitions/means	**97.81**	59.42
Transitions/variances	97.78	57.73
No randomization (baseline)	96.56	54.77

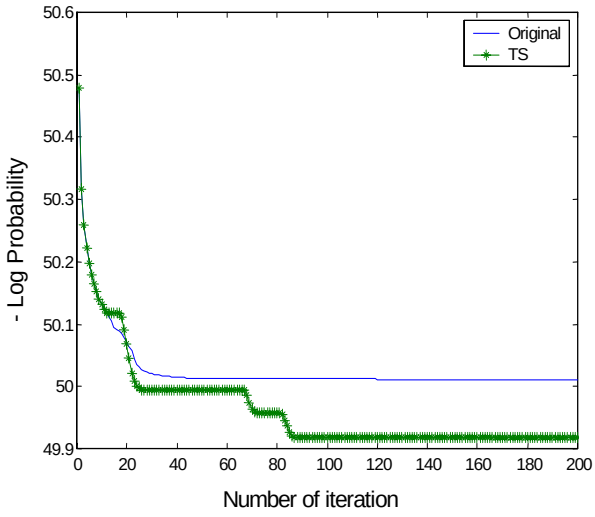

Fig.5. best_globel of normal BW and HMM with TS (Transitions/means)

From the experiment, transitions and variances don't signal any different result because random factor of these parameters are too small. There are also experiments with higher random factor, which are not presented in this paper. The experiment shows worse results with high random factor. This is because high random factor generates out-of-area parameters and out-of-area means and variances make transition toward 0. Therefore, means and variances are not good to be perturbed. However, mean perturbation is good for moving solutions out of local optima under BC.

B. SI Continuous Speech Recognition

For SI system, 42 speakers are divided into 34 speakers (17 males and 17 females) for training and 8 speakers (4 males and 4 females) for testing. The sentences are also divided into training and testing sets (376 for training and 14 for testing). Read sentences and speakers in the training set are not in the testing set and vice versa.

From last subsection, the best parameters for randomization are "transitions/means". AM in this section are extended to be multi-Gaussian CD model. Thus, mixture factor parameters are included. The experiment starts from training 8-mixture CD models (6074 tied states) and set them as an initial solution for HMM with TS.

Note that only testing with LM is performed in this section as Table 1 convinces that testing with LM can also be used for analyzing. There are 3 experiment types in this step: (1) transitions/means; (2) transitions/mixture factors; (3) transitions/mixture factors and means. Table 2 expresses the test results.

Experiment types	% Accuracy
Transitions/means	79.10
Transitions/mixture factors	**79.30**
Transitions/means and mixture factors	75.94
No randomization (baseline)	77.83

TABLE II
SI CONTINUOUS SPEECH RECOGNITION RESULTS

"Transitions/mixture factors" is the best. "Transitions/means" is worse than "transitions/mixture factors" because random mean generation can somehow destroy the relationship within each mean vector. "Transitions/means and mixture factors" gives bad result as there are too many random parameters.

V. CONCLUSIONS

HMM with TS shows a great improvement in SI speech recognition system. However, random parameters and random factor have to be carefully chosen. More parameters are randomized, more HMM parameter relationships are destroyed. The best solution in this paper is "transitions/mixture factors". The experiment displays 1.47% better than the original one.

More studies about the relationship in HMM parameters will be done in the future. We believe that better optimum still can be found if we can preserve HMM parameter relationship in the perturbation process.

REFERENCES

[1] Rabiner L. and Juang B. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ.

[2] Pham, D.T. and Karaboga D. 2000. *Intelligent Optimization Techniques*. Springer., New York.

[3] D. Puangdownreong, K-N. Areerak, A. Srikaew, S. Sujijorn, and P. Totarong. 2002. *System Identification*

via Adaptive Tabu Search. IEEE International Conference on Industrial Technology, 915-920.

[4] B. Maxwell and S. Anderson. 1999. *Training Hidden Markov Models using Population-Based Learning*. In Genetic and Evolutionary Computation Conference, GECCO-99, 994.

[5] T. Jiang and S. Ma. 1996. *Geometric Primitive Extraction Using Tabu Search*. Proc. 13th International Conference on Pattern Recognition, vol.2, 266-269, Vienna, Austria.

[6] Young S., Jansen J., Odell J., Ollasen D., Woodland P. 2000. *The HTK Book (Version 3.0)*. Entropic Cambridge Research Laboratory, Cambridge, England.