Acoustic model topology optimization for large vocabulary speech recognition

Xirimo Bao[1,*], and Chunmei Ning[2]

[1]School of Computer Science and Engineering, Hohhot College for Nationalities, Hohhot, China
[2]Library of College, Hohhot College for Nationalities, Hohhot, China

Keywords: Acoustic model, Topology optimization, Genetic algorithm, Particle swarm optimization, Speech recognition.

Abstract. Acoustic model topology selection work in constructing large vocabulary speech recognition systems is being done empirically or heuristically. In this paper, we propose two improved algorithms, which are based on Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) respectively, on the basis of our previously proposed algorithms to select and optimize model topologies for small or medium vocabulary speech recognition systems. Our improved algorithms attain the goal of optimizing acoustic model topologies for large vocabulary speech recognition systems mainly through modifying the encoding schemes of our previously proposed algorithms. Experiments on the dialogue corpus of Inner Mongolia University show that, compared with the conventional acoustic model topology selection method, our newly proposed algorithms are able to bring much higher recognition performance for large vocabulary speech recognition systems by optimizing their acoustic model topologies.

1 Introduction

In constructing HMM (Hidden Markov Model) based speech recognition systems, we need to determine the number of states and the number of Gaussian kernels per state before estimating model parameters, i.e., to select model topologies for acoustic models. The acoustic model topologies have significant impacts on the recognition performance of speech recognition systems.

In the current speech recognition community, the conventional steps taken to select acoustic model topologies for speech recognition systems are as follows: first, to select the number of states for each acoustic model empirically, and then construct single-Gaussian speech recognition system; second, to increment gradually the number of Gaussian kernels per state of the single-Gaussian system, thus obtaining two-Gaussian system, three-Gaussian system, etc.; last, to select the speech recognition system whose recognition performance is the best and the acoustic model topologies of the system are also regarded as the best.

The above mentioned method possesses the following disadvantages: first, the process of searching for the optimized kernel number per state is slow and inefficient, if the process is

[*] Corresponding author: xirimo_bao@sina.com

turned into data-driven and automatic one, the searching efficiency will be improved, and other advantages will be brought; second, it will constraint the modeling ability of acoustic models to optimize the number of states and the Gaussian kernel number per state separately, thus influencing the recognition performance of speech recognition systems; third, the strategy to allocate Gaussian kernels uniformly across model states affects the modeling precision of the acoustic models, it is reasonable to allocate the numbers of Gaussian kernels according to the real needs of the model states.

The rest of the paper is organized as follows: our previous work on acoustic model topology optimization is briefly introduced in section 2; section 3 describes the improved algorithms we newly presented; the experimental setup and results are described and given in section 4; section 5 contains our concluding remarks and future work introduction; the last section, section 6 is our acknowledgement remarks.

2 Our previous work

For the purpose of overcoming the above mentioned disadvantages of the conventional acoustic model topology selection method, we proposed in our earlier work [1-4] two acoustic model topology optimization algorithms, dubbed GA-AMTO and PSO-AMTO respectively, for small or medium vocabulary speech recognition systems.

The modeling objects of GA-AMTO and PSO-AMTO are sub-words or words and the basic strategy of these algorithms is to encode the chromosome of GA (Genetic Algorithm)[5, 6] (particle position of PSO (Particle Swarm Optimization)[7, 8, 9, 10]) as the collection of acoustic model topologies of a speech recognition system, then optimize the chromosome (particle position) by evolutions of GA (iterations of PSO), finally find the optimized acoustic model topologies of the speech recognition system in the chromosome (the particle position) whose fitness value is the highest after evolutions (iterations).

The encoding scheme of GA-AMTO and PSO-AMTO is as follows: the chromosome or the particle position is composed of one dimensional array of n×N integers, where n is the number of models and N represents the allowable maximum number of states of the models, i.e., each chromosome or particle position is divided into n parts and N integers of each part correspond to the topology of one acoustic model. As far as one model is concerned, the ith integer represents the Gaussian kernel number of the ith state of the corresponding model. Each integer ranges from 0 to the allowable maximum number of kernels per state.

For other aspects of our GA-AMTO and PSO-AMTO algorithms, e.g., algorithm structure, initialization strategy, computation of fitness value, adjustment of particle positions, etc., see [1].

3 Improved GA-AMTO and PSO-AMTO algorithms

Our improved algorithms modified the GA-AMTO and PSO-AMTO algorithms mainly through the following three strategies: first, modeling objects are changed from sub-words or words to phonemes; second, the allowable maximum number of states of the models is set to fixed value 5 to match the change made in modeling objects; third, the method to automatically generate prototype files and model definition file is also redesigned to match the change made in modeling objects.

In the improved GA-AMTO and improved PSO-AMTO, the population (the swarm) contains some chromosomes (particles), and each chromosome (particle position) represents a collection of acoustic model topologies of all modeling objects of a speech recognition system. The encoding scheme of a chromosome (a particle position) is as follows:

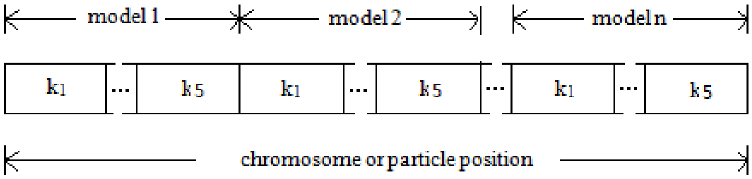


Fig.1. Encoding scheme of the improved GA-AMTO and PSO-AMTO algorithms.

As shown in Figure 1, each chromosome (particle position) is made up of 5×n integers, where n represents the number of models whose topologies are to be optimized and 5 is the allowable maximum state number for all models. Specifically, each chromosome (particle position) is composed of n parts and 5 integers of each part correspond to the model topology of an acoustic model. In a specific 5-integer part, the ith integer ki is the Gaussian kernel number for the ith state and is also one gene of the chromosome (one coordinate of the particle position in one dimension) and satisfies 0⩽ki⩽MaxKerNum, where MaxKerNum is the allowable maximum number of Gaussian kernels for each state. Allowing ki to be 0 is the key strategy taken to optimize the state number of acoustic models.

In the conventional acoustic model topology selection method, the state number of phonemes is often set to fixed value 3 empirically, so in our encoding scheme, we set the value a bit greater than 3 to include the fixed value 3 in our automatic searching space and also not to decrement the searching efficiency too much. This is the reason why we selected 5 as the allowable maximum number of states in an acoustic model in the above mentioned encoding scheme.

Because of the modification made in the encoding scheme, the process to automatically generate model definition file is also redesigned. For GA-AMTO and PSO-AMTO algorithms, we need to prepare at least 28 prototype files containing definitions of 3-state, 4-state,… , 30-state hidden markov models respectively and automatically copy the contents of the prototype files to the model definition file when needed. But for the improved GA-AMTO and the improved PSO-AMTO algorithms, we only need to prepare 2 prototype files containing definitions of 5-state and 3-state hidden markov models, where 3-state model is prepared for silence and short pause and 5-state model is prepared for other normal acoustic models.

Last but not the least, it should be pointed out that, the same as GA-AMTO and PSO-AMTO, the improved algorithms also possess the following characteristics that correspond to the above-mentioned three disadvantages of the conventional topology selection method: first, automatic and data-driven search of optimal acoustic model topologies for large vocabulary speech recognition systems using GA or PSO; second, if the number of Gaussian kernels of one state equals to zero in the optimization results, then delete the corresponding state and decrement the number of states of the acoustic model by one, thus obtaining the goal of optimizing the number of states and the number of Gaussian kernels per state simultaneously; third, in the optimization process or optimization results, the number values of Gaussian kernels per state are independent with each other, thus avoiding the disadvantage of allocating Gaussian kernels uniformly.

4 Experimental setup and results

To verify our improved GA-AMTO and PSO-AMTO algorithms, we conducted large vocabulary continuous speech recognition experiments on Inner Mongolia University speech corpus, which is made up of 1199 utterances, uttered by 46 male and 31 female speakers, and has a time length of 5 hours. The training and test corpus are composed of 1077 and 122 utterances, respectively.

The process of the experiments is as follows: we first built a context dependent multi-Gaussian large vocabulary baseline system whose acoustic model topologies are chosen using the conventional method; then, initialized the chromosomes (particle positions) of the improved GA-AMTO (the improved PSO-AMTO) algorithms with the acoustic model topologies found in the first step using the conventional method; ultimately, we repeatedly ran the improved GA-AMTO and the improved PSO-AMTO for 5 times, respectively, and compared the recognition performance of the corresponding 10 topology optimized systems with that of the baseline.

The reason why we performed the improved GA-AMTO and PSO-AMTO algorithms several times is that the optimization results of these algorithms may not be the same with each other.

The conventional method used in the experiments to select acoustic model topologies is as follows: several speech recognition systems are built, the first one is single Gaussian and the others are built by incrementing the kernel numbers of model states one by one, and then the acoustic model topologies that belongs to speech recognition system whose recognition performance is the best in the above mentioned systems are ultimately selected.

The control parameter values adopted in the verification process of the improved GA-AMTO and the improved PSO-AMTO algorithms are listed in the following table 1.

Table 1. Control Parameters of Our Algorithms.

Improved GA-AMTO		Improved PSO-AMTO	
Population Size	20	Swarm Size	20
Num. of Generaions	100	Num. of Iterations	100
MaxKerNum	30	MaxKerNum	30
Generation Gap	0.9	Vmax	8
Crossover Rate	0.7	c1,c2	2,2
Mutation Rate	0.015	wstart	0.9
Selective Pressure	2.0	wend	0.4

In the experiments, after the initialization of the chromosomes and the particle positions, the improved GA-AMTO and the improved PSO-AMTO start evolutions and iterations of acoustic model topology optimization process. The last output of these algorithms includes the optimized model topologies, word correctness rates and acoustic models of the optimized speech recognition systems.

The experimental results of the improved GA-AMTO and the improved PSO-AMTO are given in the following table 2, including WER (Word Error Rate) of ASR (Automatic Speech Recognition) system constructed on the basis of acoustic model topologies selected using the conventional method and WERs of ASR systems whose acoustic model topologies are optimized using our improved GA-AMTO and improved PSO-AMTO algorithms. As can be seen from the data in the above tables, our improved GA-AMTO and improved PSO-AMTO algorithms are both able to bring lower WER than the conventional method, so our algorithms are both more effective methods to optimize and select acoustic model topologies for large vocabulary speech recognition systems than the conventional method.

Table 2. Experimental Results of the Improved GA-AMTO and Improved PSO-AMTO algorithms.

ASR Systems		WER（%）
system optimized using the conventional method		6.30
systems optimized using the improved GA-AMTO	1	5.33
	2	5.09
	3	5.21
	4	4.97
	5	5.09
systems optimized using the improved PSO-AMTO	1	4.48
	2	4.36
	3	4.36
	4	4.24
	5	4.61

10 figures depicting model topology optimization process of the improved GA-AMTO algorithm and the improved PSO-AMTO algorithm are also plotted. They all indicate that our improved GA-AMTO and improved PSO-AMTO are two effective algorithms capable of performing acoustic model topology optimization and finding speech recognition systems with higher WCR（Word Correctness Rate）. In Figure 2 are two of the 10 figures.

In the following figures, Figure 3， are two groups of optimized acoustic model topologies obtained by the improved GA-AMTO algorithm and the improved PSO-AMTO algorithm, respectively, through their acoustic model topology optimization processes. Each group of the optimized acoustic model topologies is an array of 5 × 48=240 integers, including optimized topologies of 48 acoustic models and each optimized topology of an acoustic model is represented by 5 integers. Non-zero integer represents the optimized number of Gaussian kernels of the corresponding state and integer zero means that the corresponding model state should be removed. It can be seen from these figures that: first, the Gaussian kernels of the optimized model topologies are in many cases un-uniformly allocated; second, the maximum number of Gaussian kernels in the optimized model topologies is 25, so it is reasonable to set the allowable maximum number of Gaussian kernels to 30 in the experiments; third, there is at least one zero in each optimized model topology, so it is also reasonable to set the allowable maximum state number of each model to 5 in the experiments.

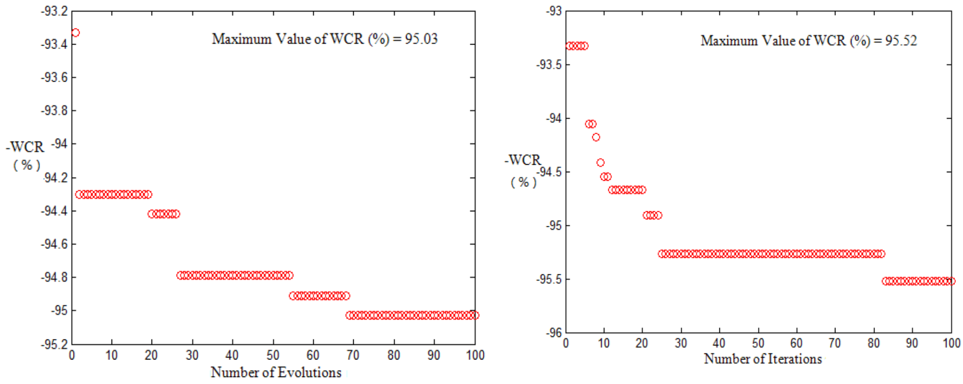
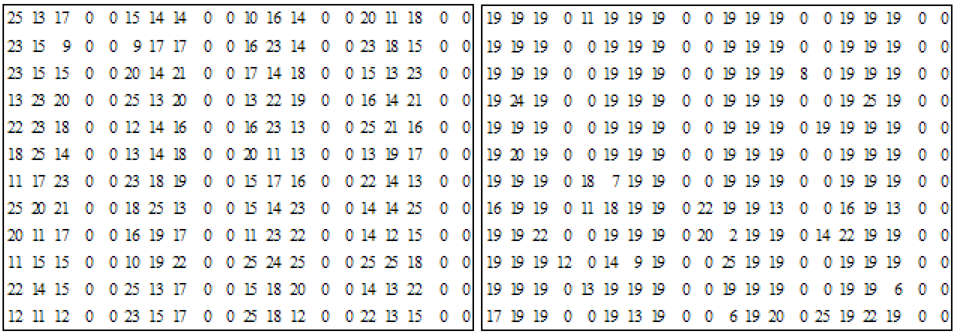


Fig. 2. Model topology optimization process of the improved GA-AMTO and PSO-AMTO.



25	13	17	0	0	15	14	14	0	0	10	16	14	0	0	20	11	18	0	0
23	15	9	0	0	9	17	17	0	0	16	23	14	0	0	23	18	15	0	0
23	15	15	0	0	20	14	21	0	0	17	14	18	0	0	15	13	23	0	0
13	23	20	0	0	25	13	20	0	0	13	22	19	0	0	16	14	21	0	0
22	23	18	0	0	12	14	16	0	0	16	23	13	0	0	25	21	16	0	0
18	25	14	0	0	13	14	18	0	0	20	11	13	0	0	13	19	17	0	0
11	17	23	0	0	23	18	19	0	0	15	17	16	0	0	22	14	13	0	0
25	20	21	0	0	18	25	13	0	0	15	14	23	0	0	14	14	25	0	0
20	11	17	0	0	16	19	17	0	0	11	23	22	0	0	14	12	15	0	0
11	15	15	0	0	10	19	22	0	0	25	24	25	0	0	25	25	18	0	0
22	14	15	0	0	25	13	17	0	0	15	18	20	0	0	14	13	22	0	0
12	11	12	0	0	23	15	17	0	0	25	18	12	0	0	22	13	15	0	0

19	19	19	0	11	19	19	19	0	0	19	19	19	0	0	19	19	19	0	0
19	19	19	0	0	19	19	19	0	0	19	19	19	0	0	19	19	19	0	0
19	19	19	0	0	19	19	19	0	0	19	19	19	8	0	19	19	19	0	0
19	24	19	0	0	19	19	19	0	0	19	19	19	0	0	19	25	19	0	0
19	19	19	0	0	19	19	19	0	0	19	19	19	0	19	19	19	19	0	0
19	20	19	0	0	19	19	19	0	0	19	19	19	0	0	19	19	19	0	0
19	19	19	0	18	7	19	19	0	0	19	19	19	0	0	19	19	19	0	0
16	19	19	0	11	18	19	19	0	22	19	19	13	0	0	16	19	13	0	0
19	19	22	0	0	19	19	19	0	20	2	19	19	0	14	22	19	19	0	0
19	19	19	12	0	14	9	19	0	0	25	19	19	0	0	19	19	19	0	0
19	19	19	0	13	19	19	19	0	0	19	19	19	0	0	19	19	6	0	0
17	19	19	0	0	19	13	19	0	0	6	19	20	0	25	19	22	19	0	0

Fig. 3. Model topologies optimized by the Improved GA-AMTO and PSO-AMTO.

5 Summary and future work

We proposed in this paper two improved algorithms, called improved GA-AMTO algorithm and improved PSO-AMTO algorithm respectively, to optimize acoustic model topologies for large vocabulary speech recognition systems on the basis of our previous work to optimize acoustic model topologies for small or medium vocabulary speech recognition systems. The above mentioned experimental results indicate that our newly presented algorithms are superior to the conventional method in selecting model topologies for large vocabulary speech recognition systems, and that our newly presented algorithms are able to find those model topologies whose Gaussian kernels are un-uniformly allocated and are also able to optimize the number of states and the Gaussian kernel number per state simultaneously.

Compared with the GA-AMTO and PSO-AMTO algorithms of our previous work, although the optimization efficiency of the improved GA-AMTO and improved PSO-AMTO algorithms has been improved a lot due to the modifications made in the encoding scheme, but they are still computationally expensive, which may constraint the further application and spread of these algorithms. Currently, if the benefits are considered, computational expenses are acceptable.

Our future work will focus on adjusting and modifying our improved GA-AMTO and improved PSO-AMTO algorithms to optimize the model topologies of deep learning [11, 12, 13, 14, 15] based speech recognition systems, including the state number of HMM, the number of hidden layers of DNN (Deep Neural Network) and the node number of each hidden layer of DNN. In addition, although the experiments to verify our improved algorithms are conducted on the Mongolian speech recognition corpus, the methods themselves are applicable to all other languages, so we will carry out verification experiments on other languages such as English and Chinese in our future work.

This research was financially supported by "Science and Technology Research Projects of Universities and Colleges of Inner Mongolia Autonomous Region", China, Grant No. NJZY17554.

References

1. X. Bao, Mongolian language oriented research on acoustic modeling for speech recognition, Inner Mongolia University, Hohhot, 2016.
2. X. Bao, G. Gao, Acoustic model topology optimization using evolutionary methods, in Conference Proceedings ACPR2011 (Beijing, China), pp.355-361, Nov.2011.
3. X. Bao, G. Gao, J. Zhang, Construction of concise speech recognition systems based on BIC and PSO, Computer Engineering and Applications, 49 (10) 14-17, 2013.

4. X. Bao, G. Gao, J. Zhang, Genetic algorithm based optimization of acoustic model topologies, Computer Engineering and Applications, 50 (14) 5-8, 2014.
5. F. Busetti, Genetic algorithms overview, Mathematical Problems in Engineering, 2015 (21) 1-9, 2001.
6. R. L. Haupt, S. E. Haupt, Practical genetic algorithms, John Wiley & Sons, 2004.
7. D. R. Umarani, V. Selvi, Particle swarm optimization– evolution, overview and applications, International Journal of Engineering Science & Technology, 38 (2) 997-1000, 2010.
8. R. Poli, J. Kennedy, T. Blackwell, Particle swarm optimization, Swarm Intelligence, 1 (1) 33-57, 2007.
9. Y. del Valle, G. K. Venayagamoorthy, S. Mohagheghi, Particle swarm optimization: basic concepts, variants and applications in power systems, IEEE Transactions on Evolutionary Computation, 12 (2) 171-195, 2008.
10. D. Sedighizadeh, E. Masehian, Particle swarm optimization methods, taxonomy and applications, International Journal of Computer Theory and Engineering, 1 (5) 486-502, 2009.
11. J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks, 61 (10) 85-117, 2015.
12. J. Pan, C. Liu, Z. Wang, Investigation of Deep Neural Networks (DNN) for large vocabulary continuous speech recognition: why DNN Surpasses GMMs in acoustic modeling, in Conference Proceedings 8th International Symposium on Chinese Spoken Language Processing, pp. 301-305, 2012.
13. T. N. Sainath, A. Mohamed, B. Kingsbury, Deep convolutional neural networks for LVCSR, in Conference Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614-8618, 2013.
14. D. Yu, L. Deng, Automatic speech recognition: a deep learning approach, Springer Publishing Company Incorporated, New York, 2014.
15. L. Deng, D. Yu, Deep learning: methods and applications, Foundations & Trends in Signal Processing, 7 (3) 197-387, 2013.