HMM speech recognition study of an Improved Particle Swarm Optimization Based on Self-Adaptive Escape (AEPSO)

Ye Liangpan1,a, He Tao2*
1 School of Automation & Electrical Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, 730070,China
2Gansu Research Center of Automation Engineering Technology for Industry&Transportation, Lanzhou Jiaotong University, Lanzhou, Gansu, 730070, China

a0618424@stu.lzjtu.edu.com
*Corresponding author’s e-mail: 837946339@qq.com

Abstract. Aiming at the single evaluation content of the current train simulation training system, the process of voice interaction in the daily work of train conductors cannot be effectively and comprehensively reflected, and the coordination between dispatching terminology and train running equipment are not considered. In order to realize the objective evaluation of whether the terms are clear and the semantics accurate and the consistency check of "schedule terms - equipment operation", the HMM speech recognition subsystem of an Improved Particle Swarm Optimization Based on Self-Adaptive Escape (AEPSO) for railway dispatching was constructed. Compared with the traditional Baum-Welch algorithm, the experiment result shows that AEPSO-BW algorithm has certain advantages in speech recognition compared with traditional Baum-Welch algorithm. It provides a more comprehensive evaluation index for the training and assessment of vehicle service personnel.

1. Introduction
In recent years, China has entered a period of deepening industrialization and urbanization, achieved major breakthroughs in high-speed railway technology, and quickly ranked first in the world in terms of the scale of high-speed railway construction and operation. The demand for railway personnel is increasing year by year. Railway companies and relevant professional colleges and universities mainly adopt centralized teaching and equipment operation in the teaching assessment of station attendants. The lack of practical training leads to the difficulty in meeting the requirements of on-site work with professional knowledge.

In train operation, the station attendant need voice interaction with each post, but at present in view of the station attendant training drill mainly depends on the traffic simulation training system. The evaluation of training personnel ,lack of voice interaction function, mainly focus on the field equipment, which can't reflect the station attendant, scheduling, public works, electricity and driver communication between the type of work coordination. Automatic Speech Recognition (ASR) technology is a technology for machines to "understand" human language. By using this technology, Speech signals issued by people are directly converted into text[1]. On the one hand, the introduction of this technology is helpful to improve the reality of the training system. On the other hand, in the learning and assessment process, there is no need for additional personnel to cooperate with the
operation and evaluate the results, so as to reduce the dependence on the assessment personnel and improve the assessment efficiency.

This study introduce an improved adaptive particle swarm optimization (AEPSO) \cite{2} HMM. The method combined with the characteristics of the species in the biological world that the living density was too high and the species would automatically move apart. By limiting the flight speed of particles in the search space and adopting adaptive strategy, it increases the retention time of particles in the search space and improves the activity degree and global search ability of particles. On this basis, the HMM parameter training method is optimized for railway scheduling speech recognition and the feasibility of this optimization method is verified.

2. Rationale

2.1. Hidden Markov model

The hidden Markov model is a statistical model for the construction of speech signal time series structure, which is considered as a mathematical double random process: one is a hidden random process that uses the Markov chain with a finite number of states to simulate the changes of speech signal statistical characteristics, and the hidden state sequence is usually composed of context-dependent three-phoneme model \cite{3}. The other is a random process of external observable Coefficients associated with each hidden state of the Markov chain. The observable states are usually composed of feature vectors extracted from the speech signal spectrum by Mel frequency Cepstral Coefficients (MFCC) \cite{4}. The hidden Markov model \cite{5} is represented by a ternary symbol $\lambda = (\pi, A, B)$. $\pi$ is an initial state probability vector, $A$ is a state transition matrix, and $B$ is an observation probability matrix. When applying HMM to practical problems, three basic problems need to be solved: evaluation problem, parameter learning problem and decoding problem. Traditional solutions are forward-backward algorithm, EM algorithm and their specific applications in speech recognition: Baum-Welch algorithm and Viterbi algorithm.

2.2. Improved Particle Swarm Optimization Based on Self-Adaptive Escape (AEPSO)

In the particle swarm optimization algorithm, the set composed of multiple possible solutions is called the population, and each individual in the population, namely the possible solution, is regarded as a particle position $p$ in the N-dimensional search space. The advantages and disadvantages of the position $p$ are determined by the fitness function. PSO generates a swarm of particles randomly first. Then, in the process of searching space flight, particles iteratively adjust the flight attitude by referring to the historical best position $p_{best}$ experienced by the individual, the historical best position $g_{best}$ experienced by all particles in the group and their own flight experience. The flight speed and position of each particle can be adjusted according to the following formula:

$$ v(t+1) = \omega \times v_{id}(t) + c_1 \times r_1 \times \left[ p_{id}(t) - x_{id}(t) \right] + c_2 \times r_2 \times \left[ p_{gd}(t) - x_{gd}(t) \right] \tag{1} $$

$$ x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{2} $$

$\omega$ is the inertia factor, $\omega \in (0, 1)$, and the global optimization ability is positively correlated with it, the local optimization ability is negatively correlated with it, $c_1$, $c_2$ is the acceleration factor, $c_1 = c_2 = 1$, $r_1$, $r_2 \in \text{rand}[0, 1]$.

In order to improve the optimization performance of PSO, an adaptive adjustment optimization strategy was introduced in combination with the characteristics that species in the biological world would automatically separate and migrate when they found that the living density was too high \cite{2}. The specific formula is as follows:

$$ \text{If} \left(v_{id} < T_d\right) \text{ then } v_{id} = r_3 \times V_{\text{max}}, \tag{3} $$
If \( v_{id} < T_d \) then \( p_{id} = x_{id} \),

\[
F_d(t) = F_d(t-1) + \sum_{i=1}^{\text{size}} b_{id}(t),
\]

\[
b_{id}(t) = \begin{cases} 0, & v_{id} > T_d \\ 1, & v_{id} < T_d \end{cases},
\]

If \( F_d(t) > k_1 \) then \( F_d(t) = 0 \), \( T_d = T_d / k_2 \).

\( T_d \) is the \( d \)-dimensional threshold velocity of the particle swarm and \( T_d > 0 \), \( r_3 \in \text{rand}[0,1] \), \( F_d(t) \) records the number of times that the particles in the particle swarm pass the \( d \)-dimensional velocity threshold \( T_d \). \( k_1 \) is the conditional value used to adjust the velocity threshold, \( k_2 \) determines the threshold velocity decline.

The algorithm randomly adjusts the particle trajectory by directional and timed mutation operations, that is, when the particle velocity is less than the current threshold velocity, the particle escape motion is given. In the process of escaping motion, particles forget their best historical position and replace the current position of particles to improve the convergence ability of the algorithm. The threshold speed affects the performance of the algorithm, so the relationship between global search and local search can be coordinated by down-regulating the threshold speed step by step.

**2.3. Initial parameter optimization of HMM**

![Diagram](image)
Based on the above analysis, this paper proposes an improved adaptive Escape particle swarm optimization (AEPSO) method for railway scheduling speech recognition to optimize the initial parameters of HMM. The initial parameter optimization and model training process of HMM are shown in Figure 1.

3. The application of speech recognition in the field of railway Vehicle Simulation and training

3.1. Virtual simulation experiment platform for railway locomotive affairs

In the process of receiving and sending trains in railway stations, the computer interlocking system, CTC system and other related equipment are all safety equipment for all-weather operation, which cannot meet the daily practical training and drill of vehicle attendants. The application of the virtual simulation experiment platform for railway traffic affairs effectively solves this problem, which is of great significance to improve the work efficiency and emergency response ability of front-line traffic affairs personnel. At the same time, the railway transport authorities can also grasp and judge the operational level of front-line traffic personnel through the platform.

This platform simulates the dialogue scene between relevant operators in the process of driving, and using voice recognition technology to detect whether the language used in object scheduling is accurate and clear. For example, the current models for dispatcher training, for driving the dispatcher training objects, training control module distributed to the various simulation module of various types of train running scene, guide the crane operator to send out the corresponding instruction for each scenario, mainly through digital dispatcher station operating interface and phonetic acquisition/synthetic voice acquisition module to perform the operation and dispatching. The speech processing module will convert the collected speech signals into text information and send it to the training control module for comparison with the scheduling corpus. If the comparison of the scheduling terms is successful, the system will record and score them, and then enter the virtual dialogue. The system collects the examinee's voice signals through the speech acquisition/synthesis module and completes the dialogue by making up the voice signals of drivers, station watchmen and maintenance personnel.

3.2. Railway dispatching voice features

The usage of railway dispatching terminology must meet the requirements of the call response standard, which mainly includes the following points: 1. The dispatching terminology is Chinese text and the pronunciation is limited to Mandarin. 2. The dialogue among vehicle service personnel is strictly in accordance with the call response standard and most of them are phrases less than 20 words. 3. The dispatch language contains a large number of professional terms and proper nouns, such as: guide pickup, switch, block, shred, "station name" and so on. 4. There are special pronunciation of Numbers or letters in railway dispatch voice. For example, "IG" is pronounced as "yigu", "K/T/Z XXX" is pronounced as "kuai/te/zhi XXX train", "1", "0", "7" and "2" are pronounced as "yao", "dong", "guai" and "liang", etc.

3.3. Network diagram of "road driving" syntactical control

Scheduling voice as a control instruction is often not independent, it needs to be used in conjunction with related scheduling equipment. The station attendant shall press the corresponding button on the operation interface of the digital dispatch desk when dictating the dispatch terms.

![Network diagram of "road driving" syntactical control.](image-url)
Considering that the importance of each word in scheduling terminology varies, it is divided into core words and general words. The core words are important information related to the traffic safety and scheduling semantics, which are directly related to the operation of the dispatching station. The general words refer to the general information which does not affect the traffic safety and the semantics of dispatching, and have no correlation with the operation of dispatching station. In Figure 2, the core words are expressed in the form of underlining, while the general words are the voluntary addition of mood auxiliaries such as "of/of/um".

4. Analysis and discussion of experimental results
This experiment uses WIN10 operating system, storage space of 1TB, sound card, microphone and other hardware platforms. Software platforms such as original sound recording and processing software Audacity, programming software PyCharm, voice testing software HTK and virtual simulation experiment platform for railway Transportation affairs.

In order to make the experimental samples meet the special requirements of railway operation and practical training assessment, the actual situation that the front-line traffic personnel are mainly male is taken into account. In this study, 5 speakers (4 males and 1 female) who meet the standard of Grade 2A of Putonghua are selected to record the experimental sample data respectively. HTK's built-in HSGen tool is used to generate the corpus, and then the statements in the corpus are recorded. Wav audio format was adopted for the sample, which included about 400 training sentences and 100 test sentences.

Conclusion The evaluation identifies the previously recorded test data with the HVite tool and calculates the replacement, deletion, and insertion of incorrect data based on the HResults tool.

| Evaluation index /% | Training algorithm | Baum-Welch | AEPSO_BW |
|---------------------|---------------------|------------|----------|
| Word Correctness    | 94.11               | 95.04      |
| Word accuracy rate  | 89.26               | 91.50      |
| Key word accuracy rate | 92.21               | 95.13      |

Table 2. Error analysis of railway dispatching speech recognition.

| Evaluation index /% | Training algorithm | Baum-Welch | AEPSO_BW |
|---------------------|---------------------|------------|----------|
| Substitution error rate | 2.30               | 1.03      |
| Insertion error rate | 4.85               | 3.54      |
| Delete error rate   | 3.66               | 3.93      |
| The total error rate | 10.74              | 8.50      |

It can be seen from Table 1 that, compared with Baum-Welch, all indexes of AEPSO_BW were improved, especially the accuracy rate of core words was significantly increased. Table 2 shows that the substitution error rate and the insertion error rate decrease to a certain extent, while the deletion error rate rises slightly and the total error rate decreases significantly.

5. Conclusion
In this paper, the practical training of railway locomotive personnel is taken as an application scenario, and the HTK and other experimental platforms developed by The University of Cambridge are used to realize the optimization of the adaptive Escape particle Swarm optimization (AEPSO) algorithm on the traditional speech recognition model, so as to explore a comprehensive performance evaluation
solution for the training personnel. On the basis of theoretical analysis and experimental verification, it is proved that this method has certain advantages and obtains expected results, which lays a foundation for subsequent engineering application.

Science and Technology Research and Development Project of China National Railway Corporation in 2019 (N2019G017)、Gansu provincial Special Fund for the Integrated Development of Information Industry and Informatization in 2019 (2019-92-10)、Lanzhou Talent Innovation and Entrepreneurship Project in Gansu Province in 2019 (2019-RC-107)

References
[1] Yu, D., DENG, L., Yu, K., et al. (2006) Analytic practice of Deep Learning Speech Recognition. Electronics Industry Publishing, Beijing.
[2] He, R., Wang, Y.J., Wang, Q., et al. (2005) An Improved Adaptive Escape Particle Swarm optimization algorithm and Experimental Analysis. Journal of Software, 12: 2036-2044.
[3] Hwang, M.Y., Huang, X.D. and Alleva, F.A. (1993) Predicting unseen triphones with senones. In: 1993 IEEE International Conference on Acoustics. Minneapolis. pp. 311-314.
[4] Gupta, K., Gupta, D. (2016) An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. In: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). Noida. pp. 493-497.
[5] Li, H. (2019) Statistical Learning Methods. Tsinghua University Publishing, Beijing.