Research on the Introduction to Models used in Speech Recognition

Xinsheng Wu[1,*]

[1]*Xinsheng School of Electrical Engineering, Shenyang University of Technology, Shenyang, China, 110870*
**Corresponding author. Email: WXS594177120@163.com*

ABSTRACT
The theoretical basis of Artificial Intelligence (AI) has been established since the middle of the last century. After 1970, after John McCarthy and others established an AI laboratory at MIT, the first batch of scholars and technologies began to emerge, and AI began to flourish. path of. Today, AI is still the focus and frontier of today's science and technology. After decades of technological precipitation, AI covers a wider range of aspects, and the knowledge system is huge and increasingly complex. Therefore, this article will introduce the basic principles and models of AI from the perspective of speech recognition. By consulting the relevant literature, this article will specifically introduce the hidden Markov model and the artificial neural network model. After the summary of these papers, we found that the hidden Markov model has very good time series modeling ability, but the classification ability is not strong. The latter has excellent classification capabilities, but cannot describe time-varying information well. Therefore, this article combines the two models. After consulting the literature, I finally found that the combination of the two models is a good choice and has great potential for development.

Keywords: *Artificial Intelligence, Speech Recognition, HMM, ANN*

1. INTRODUCTION

In 1970, T.Winograd developed the human-machine dialogue system SHRDLU, which can analyze instructions, such as understanding semantics or interpreting sentences with ambiguous meanings, and complete tasks through virtual squares. So far, AI has achieved a huge success.

Today, the development of AI is changing with each passing day. Today's AI technology has gone out of the laboratory, left the chessboard, and has been deeply and extensively applied in many industries through many service scenarios such as smart customer service and smart home appliances. AI has profoundly affected people's lives in all aspects.

It is mainly divided into three categories: cognitive AI, used to deal with complexity and ambiguity. Machine learning AI, which is at the forefront of computer science, is used to find some "patterns" in big data, and use these patterns to predict results without too much human intervention. Deep learning is today's cutting-edge artificial intelligence research field. Since it is inspired by the neural network in the human brain, it can be called an Artificial Neural Network(ANN).

However, because AI machines cannot perform different tasks in any different situations like the human brain. Therefore, the development of AI technology can never be done once and for all. With the expansion of applications, new problems are constantly being raised. Complex problems in real life often require a variety of different AI technologies and models to solve. Each type of problem can be used as a research field to carry out detailed research, which makes the knowledge system and technical details of modern artificial intelligence complex and changeable.

Usually every AI model has its advantages and disadvantages. In many cases, researchers will combine them to achieve complementary purposes. Therefore, this article gives a detailed introduction to the main model of speech signal processing, one of the fields of AI, and gives some applicable occasions by consulting relevant . This article finally made a prediction for the development of speech signal processing.

2. MAIN MODELS USED IN SPEECH RECOGNITION

With the development of computer technology and digital signal processing technology, and after many

research teams have carried out a lot of research on the theoretical research and application of speech recognition, speech recognition technology has made considerable progress[1].

Since signal models can be divided into deterministic models and statistical models[2], the main methods of speech signal processing are to use Hidden Markov Models (HMM) and ANN[1]. The following article will introduce HMM and ANN models.

2.1 Hidden Markov Models

HMM is a type of statistical model whose classic theory was proposed by L.E. Baum et al. in the late 1960s and early 1970s. This model was applied to the field of speech recognition by Jenik and others in the mid-1970s[3].

Statistical models describe the statistical characteristics of signals. At present, HMM has become the most effective statistical model for describing speech signals, and has been widely used in speech recognition, spoken language comprehension, and machine translation[2].

The voice signal has the characteristics of changing with time. Using the state description method, the change characteristics of the voice signal can be described as a transition from one state to another state. For example, when a word is generated, the system will continuously transition from one state to another, and each state will produce an output until the output of a word is completed.

Assume that the transition from one state to another only occurs at discrete moments, and the probability of transition is only related to the current state. Assuming there are L states, the transition probability can be represented by the $L \times L$-dimensional matrix A.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix} \tag{1}$$

where a_{ij} is the probability of s_i transitioning to s_j, and the sum of transition probabilities from one state to each state is 1, that is, the sum of the elements of each row of matrix A is 1. There is an N-dimensional vector b_i corresponding to each state. The output of all states is represented by an $L \times N$-dimensional matrix B, and the sum of the output probabilities of each state is 1, so the sum of the elements in each row of matrix B is also 1.

For example, here is a five states Markov process with its probability matrix A:

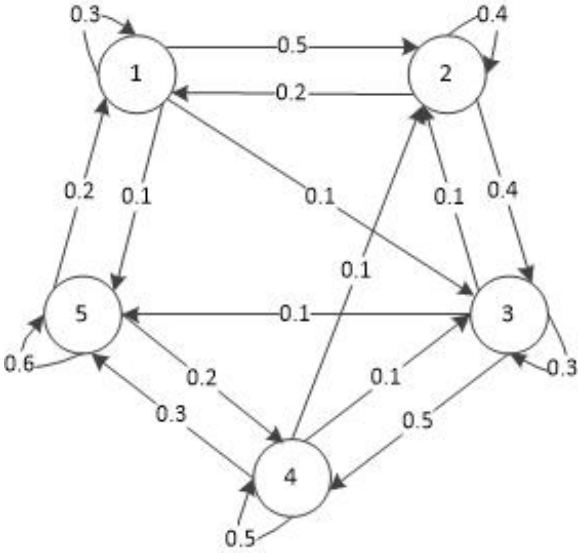

Figure 1: Five State Markov process

$$A = \begin{bmatrix} 0.3 & 0.5 & 0.1 & 0 & 0.1 \\ 0.2 & 0.4 & 0.4 & 0 & 0 \\ 0 & 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0.1 & 0.1 & 0.5 & 0.3 \\ 0.2 & 0 & 0 & 0.2 & 0.6 \end{bmatrix} \quad (2)$$

Obviously, the sum of the transition probabilities from one state to any other state is 1.

In speech recognition, the HMM model will not be as complicated as the above-mentioned example process. In the speech recognition system, the most is the HMM model from left to right. This model is more suitable for signals whose properties change over time, such as speech signals.

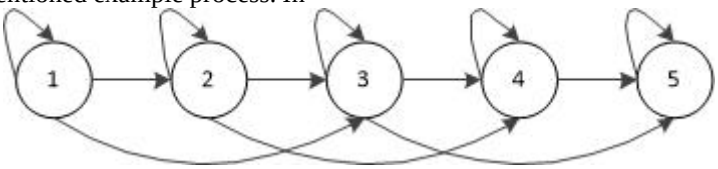

Figure 2: A left to right HMM model

In the model shown in Figure 2, the changing law of its state is: it can only repeat the original state or advance one or two states. Its complete state path is shown：

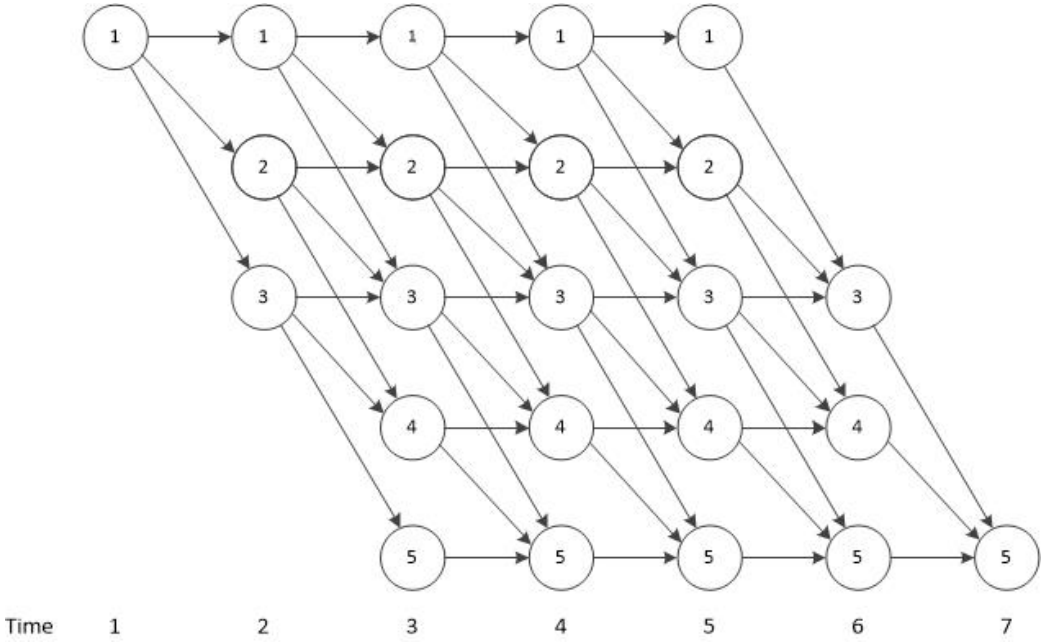

Figure 3: Seven possible path from state 1 to state 5

The main process of HMM can be divided into the following four steps[2]:

I: Select the initial state according to the initial state distribution probability π, and set the starting point n=1.

II: According to the matrix B, the output probability distribution b_{ni} in the s_i state is obtained.

III: According to the matrix A, the next state is determined by the transition probability distribution of the state transition from the state s_i at time n to the state s_j at time n+1, and n is updated so that n = n+1.

IV: If n<k , return to step II. Otherwise, the process ends.

In this way, the HMM can be defined as the following function:

$$\lambda = f(A, B, \pi) \quad (3)$$

Where π represent the probability of initial state S_j .

The HMM model can be divided into two parts: one part is a Markov chain described by π and matrix A, the other part is a random process described by matrix B, and the output is a sequence of observations. Shawn as Figure 4, where k is the length of observation time.

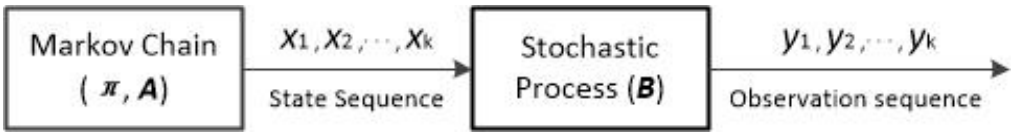

Figure 4: The composition of HMM

When HMM is used for speech recognition, three basic problems need to be solved[5] :

I: Knowing the model output Y and model $\lambda = f(A, B, \pi)$, the forward algorithm and backward algorithm are used to calculate the probability $P(Y \mid \lambda)$ of generating Y.

II: Knowing the model output Y and the model, using the Viterbi decoding algorithm to select the best observation sequence.

III: According to the model output, the model parameters are optimized to maximize the probability of matching the former, that is, to maximize $P(Y \mid \lambda)$.

The above three problems will all be encountered in speech recognition. Take isolated word recognition as an example. Assuming that there are W words to be recognized, we must first build a model for each word, divide the training sequence of each word into some states, and then study the characteristics that lead to the corresponding observation results of each state. Finally, for a given observation, find the most suitable model for it. The main process of isolated word recognition based on HMM is shown:

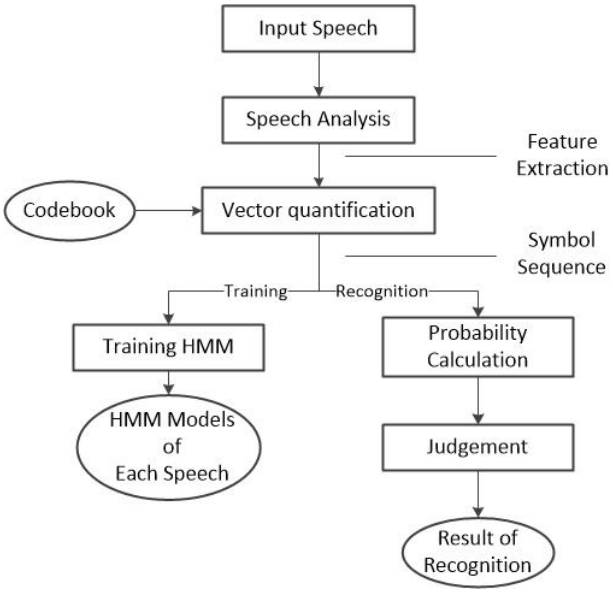

Figure 5: Isolated word recognition based on HMM

Furthermore, Hidden Markov model is able to seamlessly combine and exploit language models in the decoding process[4]. Moreover, this model has excellent theoretical and mathematical foundations and is able to robustly filter noisy information[5][6].

However, HMM also has shortcomings: in some cases, in order to reduce the amount of HMM calculations and unify the method, if the number of model states of each character sound is the same, the correspondence between the elements A and B and the physical quantities will be blurred. And when using HMM, the amount of calculation required is large and time-consuming.

2.2 Artificial Neural Network

For a long time, people have been looking forward to inventing an intelligent computer that imitates the information processing model of the human brain through the research of the human nervous system. Therefore, an artificial neural network system composed of a large number of simple processing units connected to each other has emerged.

The traditional speech signal processing system is to process the speech signal serially with symbol sequence,

which is very different from human perception. The large number of units of artificial neural network can realize distributed parallel processing, which is similar to the process of human perception and understanding of language, so it can be better applied to speech signal processing.

The Multi-Layer Perceptron(MLP) is the classic structure of neural network. It is a promotion of the Single-Layer Perceptron, which can solve the non-linear separability problem that the single-layer perceptron cannot solve. Its topology is shown in the figure below:

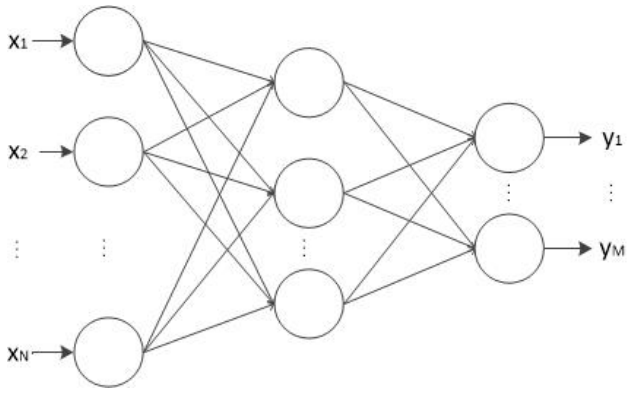

Figure 6: Topology of MLP

The MLP has one or more hidden units, and each neuron uses a differentiable activation function such as the Sigmoid Function[7]:

$$v_i = \frac{1}{1 + \exp(-\beta u_i)} \tag{4}$$

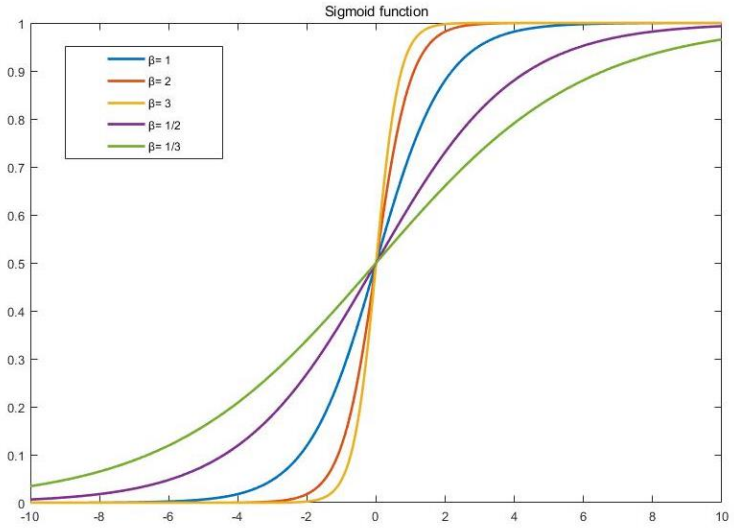

Figure 7: Sigmoid function

where u_i is input signal of a neuron, v_i is the output of the neuron. When the coefficient β is greater, the curve of Sigmoid Function will become more similar to the Step Function. As shown in figure 7.

MLP is usually trained by using Back-Propagation(BP) algorithm. The training is divided into two steps[8]:

I: Calculate the output of MLP.

II: Update the connection weight of the network with BP algorithm.

The specific method is as follows[8]:

Suppose there are N input nodes and M output nodes. First set the initial weight and threshold to a small random number, and give new input values $x_1, x_2, \cdots, x_N$, and the corresponding ideal output signals $d_1, d_2, \cdots, d_M$.

$$d_i = \begin{cases} 1, x \in i \\ 0, x \notin i \end{cases} \tag{5}$$

Where i is a certain category.

Secondly, suppose the actual output value when passing through the network are $y_1, y_2, \cdots, y_M$. For

any node j in the network, its output calculation method are

$$u_j = \sum_{i=1}^{N} w_{ij} x_i - \theta_j \tag{6}$$

$$y_j = f(u_j) = \frac{1}{1 + \exp(-u_j)} \tag{7}$$

where u_j is the sum of the weighted input and the threshold of node j , and θ_j is the threshold of node j. The non-linear transmission relationship of nodes in the network adopts Sigmoid Function.

Then modify each weight and threshold, step by step from the input node to the output layer.

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j x_i \tag{8}$$

$$\theta_j(t+1) = \theta(t) + \eta \delta_j \tag{9}$$

where $w_{ij}(t)$ is the weight from node i (output node or hidden node) to node j (hidden node or input node) at time t. x_i is the input signal on the i-th input node or the output signal on the i-th hidden node. η is the gain factor, which is a constant representing the learning rate. Usually, $0 < \eta < 1$. δ_j is the weight correction factor of node j.

When node j is the output node, the ideal output is clear, therefore δ_j can be obtained by the following formula

$$\delta_j = y_j(1 - y_j)(d_j - y_j) \tag{10}$$

When node j is a hidden node, the ideal output is not clear, so δ_j is defined as

$$\delta_j = x_j(1 - x_j)\sum_m \delta_m w_{jm} \tag{11}$$

where m is the total number of nodes in the former layer of the hidden node j.

Finally, repeat the second step until w_{ij} and θ_j are stable.

Since the BP algorithm is essentially a gradient method, there will be a local minimum point problem. The training speed of the BP algorithm is slow, and there is no theoretical guide to the selection of the number of hidden nodes.

In addition to BP Neural Network, ANN models used in speech signal processing also include RBF(Radial Basis Function), SOFM(Self-Organizing Feature Map), and TDNN(Time Delayed Neural Network). Compared with the BP network, the RBF network not only has a physiological basis, but its hidden layer nodes use a radial basis function. The output of the RBF network is the linear weighted sum of the hidden layer output, the structure is more concise, and the learning speed is faster. The output of SOFM network provides statistical estimation data for the HMM model, which is used to calculate the state transition of the HMM and the output probability of each state[2]. And TDNN can better deal with the dynamic characteristics of speech.

3. DISCUSSION OF HMM&ANN

HMM is the main method of speech recognition, which has a relatively powerful ability to model time series structure, but the recognition ability of HMM is weak. Because this model is different from the way the human brain works, its adaptive ability and robustness are not ideal. It has a poor ability to model low-level acoustic phonemes, which can easily cause word confusion. For high-level semantic understanding, it also does not have good modeling capabilities. In addition, the first-order HMM model also needs to make a priori assumptions about the state distribution.

ANN simulates the working principle of the human brain neural network, which is essentially an adaptive nonlinear kinetic system[8] with learning, association, contrast, reasoning and generality. These are not available in HMM. However, ANN also has a shortcoming that it does not describe the time dynamics of the speech signal, since the input node of the ANN is fixed, but the speech signal changes with time. Therefore ANN cannot be a main method of speech recognition., but it cannot describe the time-varying characteristics of speech signals well[2].

Therefore, the combination of HMM and ANN can make full use of the advantages of HMM's strong time regulation ability and ANN's strong resolution ability to obtain better time matching and pattern classification.

The HMM/ANN hybrid system has the following advantages: HMM parameters are obtained by ANN, and HMM does not need to make many assumptions. The model parameters calculated by ANN are related to the time-varying characteristics of the speech signal. Using ANN to calculate the speech model parameters[2][9] can make the model parameters establish the best matching relationship with this type of speech.

Since the training algorithms of HMM and ANN require a lot of calculations, they require a long

calculation time. However, as hardware systems are updated and promoted and voice recognition algorithms are improved, voice access control systems based on HMM/ANN will be more and more widely used[10].

The main limitation of the hybrid HMM/ANN system is that it is difficult to model context-dependent sub-word units[11]. Since the output of each neural network is related to the state of HMM certainty, it is necessary for ANN to model all context-related sub-word units. However, due to complexity and data sparsity issues, it is impractical to train a neural network to model all possible context-sensitive sub-word units.

On the other hand, the combination of ANN and other information processing technologies has also become an important development direction and has received more and more attention. For example, the combination of ANN and Genetic Algorithm (GA) can optimize the network topology of ANN and improve the effective learning algorithm[2].

Nowadays, the research work of speech recognition tends to solve the practical problems faced in real environment applications[12]: for example, the study of spoken speech recognition in real scenes with multiple people and multiple parties is one of the research focus of speech signal processing. Another trend of current speech recognition research is to no longer only focus on the accuracy of continuous speech in the large vocabulary. Second, from the perspective of practical applications, actively explore ways and methods for machines to perceive and understand human speech[12].

4. CONCLUSION

Through the analysis of this article, it is found that the HMM has a relatively strong modeling ability for time series structure, but its recognition ability is not strong, which is easy to cause word confusion, and its robustness is not ideal. The ANN model simulates the working mode of the human brain and has powerful learning, reasoning and classification capabilities for speech recognition, but it cannot describe the time-varying characteristics of speech signals, so it cannot be used as a mainstream speech recognition method.

From this we conclude that using ANN to give HMM parameters can combine the two models, and save the complicated assumptions of HMM. The HMM/ANN hybrid model combines the advantages of the two models, can establish the best matching relationship between the model parameters and the speech, and has great development potential.

REFERENCES

[1]Zhu Xuefang, Xu Jianping. Computer speech signal processing and speech recognition system [J]. Journal of Nanjing University of Posts and Telecommunications, 1998(Z1): 117-123.

[2]Hu Hang. Modern Speech Signal Processing. Beijing: Publishing House of Electronics Industry, 2014.

[3]Wang Chunling. The establishment and application of hidden Markov model [D]. National University of Defense Technology, 2002.

[4]Marino Maria Francesca, Alfo Marco, Gaussian quadrature approximations in mixed hidden Markov models for longitudinal data:A simulation study, Computational Statistics & Data Analysis, 2016,94: 193-209.

[5]Baumgartner Josef, Georgina Flesia Ana, Gimenez Javier, Pucheta Julian, A new image segmentation framework based on twodimensional hidden Markov models, Integrated Computer-aidedEngineering, 2016, 23(1): 1-13.

[6]Han Gain, Sohn Keemin, Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model, Transportation Research Part B-methodological, 2016, 83: 121-135.

[7]Gu Yanchun. Matlab R2016a 27 cases of neural network design and application. Beijing: Publishing House of Electronics Industry,2018.

[8]Zhang Xueying. Digital speech processing and Matlab simulation. Beijing: Publishing House of Electronics Industry, 2016.

[9]G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[10]Li Bo, Wang Dong-xia, Zou De-jun and Hu Tie-sen, "On speech recognition access control system based on HMM/ANN," 2010 3rd International Conference on Computer Science and Information Technology, 2010, pp. 682-686, doi: 10.1109/ICCSIT.2010.5565135.

[11]M. Razavi, R. Rasipuram and M. Magimai-Doss, "On modeling context-dependent clustered states: Comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 7659-7663, doi: 10.1109/ICASSP.2014.6855090.

[12]Han Jiqing, Zhang Lei, Zheng Tieran. Speech Signal Processing(Third Edition). Beijing: Tsinghua University Press, 2019.