The Deep Neural Network and Content Transcription-based Speech Recognition Algorithm in Keyword Detection

Lei Zheng
School of Information Engineering, Shandong Youth University Of Political Science, Jinan, China
87137580@qq.com

ABSTRACT Objective: It is hoped that the DNN-Hidden Markov Model (HMM) acoustic model and the long-short-term memory neural network-Hidden Markov (LSTM-HMM) acoustic model can be used to improve the accuracy rate of speech recognition in keyword detection. Method: First, the principles of speech recognition and related algorithms are introduced. Then, a DNN algorithm is applied to the speech recognition system, and a keyword detection system is established based on the DNN-HMM acoustic model. As for the experimental comparison, the proposed DNN model is used for acoustic modeling, and the effect of the DNN algorithm on the performance of the recognition system is analyzed experimentally. Results: The size of the training parameters of the proposed LSTM model is 436570, the size of the training parameters of the deep neural network (DNN) is 698100, and the size of the training parameters of the Gaussian mixture model (GMM) is 1226700. The accuracy of speech recognition based on the LSTM-HMM model and DNN-HMM model is 96.5% and 91.6%, respectively, which is significantly higher than 78.5% of the traditional speech recognition model (GMM-HMM). Conclusion: The speech recognition technology based on LSTM-HMM and DNN-HMM models has higher accuracy and is more suitable for speech keyword detection.

1. INTRODUCTION
With the coming of the digital age, in the current explosive environment of information growth, the traditional way of saving information in text form has been unable to meet the needs for knowledge of modern people [1]. Voice, as a direct medium for information recording and dissemination, can impart an emotion to information while transmitting information in real-time, which is more valuable for information recording [2]. The importance of voice-carrying information has become increasingly prominent, and almost all multimedia files contain voice data. With the large increase in the application of multimedia files, information retrieval technology based on multimedia data has become the focus of research in information science [3-5]. The means to find the information of interest from various multimedia documents as quickly and accurately as retrieving the text has been a hot topic that has received widespread attention at the moment.

This study introduces the principle of speech recognition and related algorithms. Based on these principles, a DNN algorithm is applied to the continuous recognition system of speech volume, and a keyword detection system is established based on the DNN-HMM acoustic model. As for the experimental comparison, the proposed DNN model is used for acoustic modeling, and the effect of the DNN algorithm on the performance of the recognition system is analyzed experimentally.
2. Related Work
Keyword detection technology originated in the 1970s, and the earliest research was based on the concept of “given word”. As a key technology in keyword retrieval, speech recognition has received widespread attention[6]. In 2006, Hinton et al. proposed the concept of deep learning. Microsoft researchers introduced the restricted Boltzmann machine (REM) and deep belief network (DBN) into the speech recognition acoustic model training, which was successful among large vocabulary speech recognition systems[7].

The speech recognition research in China started late. With the strong support of the country, the Chinese Academy of Sciences Institute of Automation, the Chinese Academy of Sciences Institute of Acoustics, and other scientific research institutions have undergone intensive research on speech recognition and have made significant progress[8]. At present, Microsoft, IBM, Google, and other foreign companies have successively developed Chinese speech recognition systems[9]. Chinese companies such as Xunfei, Baidu, and Sogou have also launched corresponding products in the Chinese continuous speech recognition project. In summary, the speech recognition technology in the keyword detection system has a very broad development prospect in the future. However, speech recognition technology still faces various challenges and cannot be effectively avoided in terms of speech recognition errors[10]. Therefore, this study wishes to study the speech recognition algorithm based on DNN, hoping to provide new ideas for improving the performance of speech recognition in the speech keyword retrieval system.

3. Methods

3.1. Flows and principles of speech recognition
A complete speech recognition system includes functions of speech preprocessing, speech feature extraction, speech model library construction, and speech pattern matching. For the recorded speech signal, the speech pre-processing operation is first performed. The pre-processing consists of steps such as sampling, quantization, filtering, pre-emphasis, windowed framing, and endpoint detection. The next step is to perform speech signal feature extraction. The purpose is to extract the feature parameters that can characterize the nature of the speech signal, as well as remove the irrelevant noise signals to obtain input parameters that can be used for acoustic models or speech recognition. The structure of speech recognition and speech preprocessing flows are shown in Figure 1.

![Figure 1: Speech recognition structure and speech preprocessing process](image)

3.2. Algorithms and model basis for speech recognition
The core of speech recognition is its acoustic model. Currently, the HMM is mainly used to model the temporal change of speech signals. The observation probability estimation methods in each state of the HMM can be divided into discrete, semi-continuous, and continuous types. At present, speech recognition systems are mainly continuous or semi-continuous. When describing the acoustic layer model through HMM, the hidden state corresponds to the relatively stable pronunciation state of the acoustic layer, which can describe the dynamic changes of the speech signal.
There are 6 states of the HMM model in Figure 2, 4 of which are launching states. The leftmost state indicates the starting state. Each hidden state will launch a state outward according to the probability distribution and then turn to a state to the right. The rightmost end-state indicates that the HMM has ended. At a certain time node t, the model has a sequence of states. At t+1, each state of the model will turn to a new state, representing a new sequence of states. T is not final until it reaches the final moment. One of the most important features of this process is that the state at time t is only related to the state at time t-1, which is known as Markov. The basic composition of HMM includes:

1. The state set \( S = \{s_1, s_2, ..., s_N\} \), where N represents the number of phonemes. Due to the need to consider the influence of context in speech recognition, a three-phone model is often used;
2. The state transition matrix A; and
3. The output distribution \( B = \{b_j(x)\} \), indicating the initial probability of each state.

### 3.3. Combination of DNN and traditional acoustic models

The current keyword detection systems are usually based on Large Vocabulary Continuous Speech Recognizers (LVCSR). In the speech keyword retrieval system, the GMM-HMM model combining GMM and HMM is traditionally used as the acoustic model of LVCSR, but it has a low recognition rate for speech signals. With the huge impact of deep learning technology in the field of speech recognition, the use of DNN instead of GMM to form DNN-HMM acoustic model has attracted wide attention. The DNN model is a feed-forward neural network model with multiple hidden layers. The DNN model has a total of L+1 layers, where layer 0 is the input layer, layers 1 to L-1 are hidden layers, and layer L is the output layer. The adjacent layers are connected by a feed-forward weight matrix.

In most cases, the activation function of the DNN model is the sigmoid function:

\[
\sigma(z) = \frac{1}{1 + e^{-z}}
\]  

(1)

The output range of \( \sigma(z) \) is (0,1), which helps to obtain a sparse expression. However, it makes the activation value asymmetric. For multi-classification tasks, each output neuron represents a class of \( i \in \{1, 2, ..., C\} \), where \( C = N_L \) is the number of classes. Given training criteria, the model parameter \( C = N_L \) can be learned by using the well-known error back-propagation algorithm, and the chain rules can be used for derivation. In its simplest form, the model parameters are optimized by using the first derivative information according to the following equation:

\[
W_{l+1}^i \leftarrow W_l^i - \varepsilon \Delta W_l^i
\]

(2)

\[
b_{l+1}^i \leftarrow b_{l}^i - \varepsilon \Delta b_{l}^i
\]

(3)

Where: \( W_l^i \) and \( b_l^i \) are the weight matrix and bias vector of the l layer after the t-th iteration update, respectively.
\[
\Delta W_t^l = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla J(W, b; o^m, y^m) \\
\Delta b_t^l = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla J(W, b; o^m, y^m)
\]

Equations (4) and (5) are the average weight matrix gradient and average bias vector gradient obtained after the t-th iteration, respectively. These are obtained using \( M_b \) pieces of training samples, where \( \epsilon \) is the learning rate, and \( \nabla_x J \) is the gradient of \( J \) relative to \( X \).

The model parameters of the DNN need to be trained by the training sample \( S = \{ (o^m, y^m) \mid 0 \leq m \leq M \} \) for each task. Where: \( M \) is the number of training samples, \( o^m \) is the m-th observation vector, and \( y^m \) is the corresponding output vector. This process is called the training process, or the parameter estimation process. It needs to be given a training criterion and a learning algorithm. In speech recognition tasks, acoustic model training is used to complete this process. For a fully connected DNN between adjacent layers, the weight is initialized to a small random value to avoid all hidden units in a layer getting the same gradient. DNN with many hidden layers is difficult to optimize. DNN may need to be extended beyond the test dataset.

Since speech signals are time-series signals, DNN cannot directly model them. HMM is used to model the dynamic change of the speech signal, and the observation probability is estimated by DNN. The structure of the DNN-HMM model is shown in Figure 3.

The main training steps of DNN-HMM are: (1) aligning the training set with a conventionally trained DNN-HMM model to obtain alignment information; (2) building a mapping of context-sensitive state to speech ID; (3) generating the information based on input and output labels required for training DNN; (4) obtaining HMM model structure required in DNN based on mapping relationship and GMM-HMM model; (5) estimating the prior probability of speech based on input and output labels; and (6) adjusting the network parameters by using the back propagation algorithm to obtain the DNN-HMM model.

![Figure 3: The structure of DNN-HMM model](image)

4. Results

4.1. Experimental data
This experiment chose an open-source Chinese Mandarin speech database --aishell, which is provided by Beijing Shell Company. The voice materials in the database were performed on the test set of the same speaker. Under a quiet environment, computer voice recording software Cool Edit Pro was used to record voice messages. Eight speakers were chosen to read 20 education words in turn, 10 times for each word. The sampling frequency was set to 8 KHz; each sampling point was 16 bits quantized and
stored in the mono track. A total of 1600 voice samples were obtained as training and recognition corpus. The first 3 passes and the last 3 passes were used as the training set, with a total of 960 samples; the middle 4 samples were used as the same speaker test set, with a total of 640 samples.

In the feature extraction of the speech signals, the 24-dimensional Mel-frequency Cepstrum (MFC) coefficient feature was extracted from each pre-processed speech sample of the training set and the same speaker test set, which was regularized by the mean value variance. The feature had a 25 ms window size and 10 ms overlap. The speech recognition performance of the traditional ANN model and the DNN model proposed in this study was compared. The accuracy rate of speech recognition was used as the evaluation standard. The values were statistical average results.

4.2. Construction of speech keyword retrieval system based on speech recognition

The speech keyword retrieval system includes two parts, i.e., indexing and keyword retrieval. Among them, the indexing consists of three parts, i.e., speech recognition, post-processing of speech recognition, and index construction. Keyword retrieval consists of two parts, i.e., the key inspection and confidence evaluation, as shown in Figure 4. The speech recognition errors and foreign words greatly affect the recall rate of the system. The fuzzy matching method can improve the recall rate effectively; however, at the same time, it increases the query time. During a keyword query, an initial quick lookup can be performed in the superclass database to narrow down the search, and then an exact lookup can be performed in the syllable sequence database to speed up the search time.

![Figure 4: Components of a speech keyword retrieval system](image)

The speech keyword retrieval system depends on the recognition results. Therefore, the performance of speech recognition has a vital impact on the retrieval performance of the system. In general, the performance of speech recognition systems is evaluated by the recognition error rate and real-time rate. In a speech keyword retrieval system, the recognition process of speech data can be performed offline, so that the real-time indicators are not considered. In the recognition results, the types of errors include insertion, deletion, and substitution errors. The optimal result of recognition is compared with the reference text, and the recognition error rate can be obtained.

4.3. Comparison of speech recognition results

The frame number of the feature parameters of the speech signal is set to 23. The non-linear tanh function is selected as the activation function. The output is 30 neurons, which makes the number of output neurons the same as the number to be classified. The cross-entropy between the estimated probability distribution and the actual probability distribution is used as the objective function, and the iteration is stopped when the accuracy of speech recognition improves below 0.2%. The recognition accuracy results of different speech recognition algorithms are shown in Table 1.

| Model         | Recognition accuracy |
|---------------|----------------------|
| GMM-HMM       | 78.5                 |
| DNN-HMM       | 91.6                 |
| LSTM-HMM      | 96.5                 |

As shown in Table 1, the accuracy rate of speech recognition based on the LSTM-HMM and DNN-HMM models is significantly higher than that of the traditional GMM-HMM model. Meanwhile, the
speech recognition accuracy rate of the LSTM-HMM model can reach 96.5%. Therefore, the proposed model has better performance. This study also analyzes the training parameter size of different speech recognition algorithms. The LSTM training parameter size is 436570, the DNN training parameter size is 698100, and the GMM training parameter size is 1226700. In the case of limited speech samples in the training set, the over-fitting of the training model will cause the problem of over-fitting of the training model. Therefore, speech recognition based on DNN can reduce the size of the training parameters and avoid over-fitting of the training model effectively.

The speech signal has strong randomness, and the number of frames of the speech feature parameters extracted by the same pronunciation unit may be different. The effect of the regular frame number on the recognition performance of different algorithms is shown in Figure 5. As the number of regular frames increases, the input gets closer and closer to the original feature parameters, and the recognition accuracy of the two network models continues to increase. The model training process is achieved by calculating the mean square error through the stochastic gradient descent method and then adjusting the network parameters to reduce the mean square error. Therefore, the convergence of the network model directly reflects whether the overall performance is superior.

![Figure 5: Effect of regular frame number on recognition performance of different algorithms](image)

5. Conclusions

To solve the problem of the low recognition rate of the GMM-HMM acoustic model applied in the traditional keyword detection system, this study applies a DNN-based speech recognition algorithm to the keyword detection task. The DNN-HMM acoustic model is mainly used to replace the GMM-HMM model in the original system, and the keyword detection is explored on this basis. The comparison experiment in this study chose an open-source Chinese Mandarin speech database --aishell. It was performed on the test set of the same speaker. Under a quiet environment, computer voice recording software Cool Edit Pro was used to record the voice information. Through comparison, the accuracy of speech recognition based on the LSTM-HMM model and DNN-HMM model is 96.5% and 91.6%, respectively, which is significantly higher than 78.5% of the GMM-HMM. This proves that the performance of the LSTM-HMM model proposed in this study is better. In the case where the training set has limited speech samples, the training parameter scale is too large and the training model is overfitting. Therefore, the speech recognition algorithm based on DNN can reduce the scale of training parameters, thereby effectively avoiding the problem of overfitting the training model.

The LSTM-HMM-based speech recognition technology has higher accuracy and is more suitable for speech keyword retrieval. In addition, the DNN-based speech recognition algorithms can reduce
the scale of training parameters, thereby avoiding the overfitting problems of the training model effectively. In the context of complex speech, the robustness of keyword detection still has great potential for improvement. Therefore, the subsequent research can explore the direction of extracting more robust acoustic features, and accurately retrieve the required speech information in the case of noise interference.

Acknowledgments
This work was supported in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J16LN12.

REFERENCES
[1] Shah, M., Arunachalam, S., Wang, J., et al. (2018) A fixed-point neural network architecture for speech applications on resource constrained hardware. Journal of Signal Processing Systems, 90(5), pp.27-741.
[2] Fang, S. H., Tsao, Y., Hsiao, M. J., et al. (2019) Detection of pathological voice using cepstrum vectors: A deep learning approach. Journal of Voice, 33(5), pp.634-641.
[3] Sainath, T. N., Weiss, R. J., Wilson, K. W., et al. (2017) Multichannel signal processing with deep neural networks for automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(5), pp.965-979.
[4] Mustafa, M. K., Allen, T., Appiah, K. (2019) A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. Neural Computing and Applications, 31(2), pp.891-899.
[5] Serizel, R., Giuliani, D. (2017) Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Natural Language Engineering, 23(3), pp.325-350.
[6] Price, M., Glass, J., Chandrakasan, A. P. (2017) A low-power speech recognizer and voice activity detector using deep neural networks. IEEE Journal of Solid-State Circuits, 53(1), pp.66-75.
[7] Perol, T., Gharbi, M., Denolle, M. (2018) Convolutional neural network for earthquake detection and location. Science Advances, 4(2), pp.e1700578.
[8] Zhang, M., Diao, M., Gao, L., et al. (2017) Neural networks for radar waveform recognition. Symmetry, 9(5), pp.75.
[9] Norman, B., Pedoia, V., Noworolski, A., et al. (2019) Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. Journal of digital imaging, 32(3), pp.471-477.
[10] Albrecht, T., Slabaugh, G., Alonso, E., et al. (2017) Deep learning for single-molecule science. Nanotechnology, 28(42), pp.423001.