Speaker identification based on deep learning in FX iDeal system

Yan Ke\(^1\text{a}\), Li Na\(^2\text{b}\), Chen Yuting\(^3\text{c}\)

\(^1\)Xi’an Jiaotong University No.28, Xianning West Road, Xi’an Shanxi, China +86-15002923953
\(^2\)CFETS Information Technology (Shanghai) Co., Ltd. Building 7,1388 Zhangdong Road, Pudong New Area, Shanghai, China +86-15102117735
\(^3\)CFETS Information Technology (Shanghai) Co., Ltd. Building 7,1388 Zhangdong Road, Pudong New Area, Shanghai, China +86-13564634556

\(\text{a}15002923953@163.com \text{ b}lina_zh@chinamoney.com.cn \text{ c}chenyuting@chinamoney.com.cn\)

ABSTRACT The FX (foreign exchange) iDeal system, which provides instant foreign exchange information for inter bank traders, with the rapidly increased number of users, there are urgently needs to meet the higher security requirements and improve the compliance of transactions in the inter bank market. In this paper, we make a research on speaker recognition to improve the security of the system. According to the historical sound data sent by the user, we extract the Mel cepstrum features, and then construct the speaker's identification model based on the deep neural network (DNN). Based on the model, we can verify the user. In order to prevent the DNN from over-fitting or falling into the local minimum, we add a dropout mechanism and a dynamic learning rate adjustment algorithm to the network. To maximize the data mining capabilities, we use the restricted Boltzmann machine (RBM) to realize unsupervised training which can make the network initialization parameters are closer to the global minimum, and then optimize the parameters by supervised training. Verification experiment shows that the model can effectively improve the speaker confirmation ability and is more effective than the general BP neural network and GMM model.

1. INTRODUCTION
Nowadays, with the continuous expansion of Chinese bank foreign exchange market, the National Inter bank Funding Center has built a foreign exchange instant messaging system named as FX iDeal, which provides a unified piece of information exchange service for all of bank traders in order to improve the market compliance. As the number of users increases rapidly, the security of the system has received widespread attention, especially in the user chat module which can involve in confidential transaction information interaction. Therefore, ensuring the login user is not confessed is very important. Traditional institutional verification and real-name authentication methods may reveal user information. Human biometrics are non-replicable and voice features are easier to collect. Therefore, the identity authentication based on voice feature has received extensive attention.

The speaker recognition is a biometric identification method which uses voice features, and its implementation is composed of two parts: speaker voice features extraction and the establishment of speaker voice feature model. In terms of voice feature extraction, commonly methods are linear
prediction coefficient (LPCC) [1] method, Mel cepstral coefficient (MFCC) [2] method, and traditional methods combined with time-frequency analysis, wavelet analysis, neural network and other extracted acoustic features. In terms of the model building, there are Gaussian mixture models - General Background Model (GMM-UBM) [3], Joint Factor Analysis (JFA) [4] model, and Total Difference Space Factor (i-vector) [5] model. However, with the arrival of big data era, traditional modeling methods are no longer able to obtain information from data effectively. In 2006, Geoffrey Hinton introduced the strategy of “greedy layer-by-layer pre-training” into the neural network [6], which can train larger and deeper networks. From then on, deep learning has received widespread attention from scholars. In 2012, Hinton applied the deep neural network (DNN) to the acoustic model of speech recognition for the first time [7], which verified that DNN had more powerful modeling capabilities and representation features than GMM in large vocabulary continuous speech recognition tasks. DNN becomes the mainstream modeling method for acoustic models [8][9]. However, few scholars use deep learning to construct the speaker recognition model.

In this essay, to improve the security of the FX iDeal system, we aim at the identification problem in the system and research on the speaker recognition technology by considering of stability of the biometric identification. By collecting the voice chat records of the user, the MFCC features can be extracted and the DNN can be built to construct the speaker recognition model. In addition, we use the Boltzmann machine [10] to pre-training the network. To improve the network performance, we add the dropout mechanism [11] and dynamic learning rate adjustment algorithm to the network. To verify the speaker confirmation ability of the DNN training model, we compare it with the GMM-UBM method and the general BP neural network method. In the end, the model will be applied to the FX iDeal system to strengthen the security of the system.

2. User authentication model based on DNN

2.1 The instant messaging system

The FX (foreign exchange) iDeal system is a software platform for various bank traders to provide instant chat services, market display services, regulatory support services, and transaction assistance services. According to the plans of the National Development Department, the associated system of the instant messaging system includes the upstream system, service invocation system, and external user system. If the transaction right is available, users can use the transaction assistance, supervision support and other functions when they log in the system using the unified terminal.

The overall architecture of the system follows three major architectures, including data architecture, application architecture, technical architecture, and six technical layers, including user access layer, access service layer, application service layer, data service layer, technology platform layer and infrastructure layer. All modules of the system include instant messaging module, login authentication module, transaction agent module, field supervision module and interface module. Among them, the instant messaging module provides chat services, including compliance supervision, data analysis, and smart recommendation. User identity authentication in the instant messaging module is one of the important channels to improve the system security. In instant messaging, token authentication is required for each communication. After the verification is successful, the cache is updated, otherwise the communication is rejected. Speaker identification is provided during a chat in the FX iDeal system.

2.2 speaker recognition model based on DNN

The complete speaker recognition system consists of two parts, the first part is the voice feature extraction, and the second part is building the speaker's identification model. Since the MFCC coefficient features are designed in consideration of the principle of human auditory perception, the relationship between the frequency and the actual frequency is shown in equation (1), and has a good characterization ability [12]. This paper extracts MFCC coefficients from user voice data.

\[
 f_{Mel} = 1127.0 \ln\left(1 + \frac{f}{700.0}\right)
\]  

(1)
Where \( f \) is the physical frequency in Hz.

The deep neural network, used to construct the speaker’s identification model, is a multi-layer neural network containing multiple hidden layers, as shown in Figure 1. The network input are the voice features. After the hidden layer network, the network output is the verification result of speaker, including two neurons.

At the input layer, the voice features are MFCC coefficient features of one frame or consecutive frames. Taking the 5-frame and 36-dimensional MFCC coefficients as an example, the input layer are 5*36 neurons.

In the training phase, first, the network parameter pre-training is performed using a restricted Boltzmann machine, which is an energy model based on the Boltzmann machine. The Boltzmann machine originated from physics, which was simplified and successfully applied to the field of machine learning by Hinton et al. After network pre-training, the network parameters are in a better position. After that, the network training is performed to optimize the parameters. To avoid over-fitting, the dropout mechanism is added which applies a kind of method to set the hidden layer nodes as 0 with a certain probability during training. In addition, in order to prevent network oscillation or slow convergence, a dynamic learning rate adjustment algorithm is added to the network, and the change of the learning rate is determined by the error between the two training sessions in the network, as shown in equation (2).

\[
\begin{cases}
\alpha \cdot r(i) & E(i) < E(i-1) \\
\beta \cdot r(i) & E(i) > E(i-1) \\
r(i) & \text{other}
\end{cases}
\]

where, \( r(i) \) represents the current learning rate, \( r(i+1) \) represents the updated learning rate, \( E(i) \) represents the current training error, and \( E(i-1) \) represents the last iteration error.

In the output layer, two neurons are included to characterize the judgment of the confirmation of the speaker. In this paper, the true speaker is represented by outputs 1 and 0, and the false speaker is represented by outputs 0 and 1.

Finally, a feasibility analysis of the network is performed, and the error rate of the model identification is calculated using the formula (3):

\[
\text{err} = \frac{N}{S}
\]

Where \( N \) represents the number of output errors, \( S \) represents the number of samples.
3. Experimental verification and analysis
The voice test data set used in this paper is the TIMIT voice library established by MIT, and is an authoritative database for speaker recognition. It contains 192 women and 483 male speakers from different parts of the United States. Each speaker has 10 voices and the data set is a total of 6300 voice data. Among them, each voice recording length is about 3-5s, and the acquisition frequency is 16kHz.

In the experiment, n speakers were randomly selected from the speech library, of which 1 was used as the speaker to be recognized, and the remaining n-1 were used as the false speaker. The 8 speech data of each speaker is used as the DNN model test set, and the remaining 2 speech data is used as the DNN model test machine. The 36-dimensional MFCC coefficient feature of each voice is extracted from each speaker's voice, and the voice data is framed by a 256-point Hamming window.

The neural network training takes a frame or consecutive frames of MFCC features as input, and sets the label (0, 1) or (1, 0) to represent the true speaker and the false speaker. In this paper, five consecutive frames are selected, with a total of 5*36=180 features as input; the hidden layer is set to three layers; and the output layer is set to two neurons according to the label. The network structure is 180-100-50-2. In this paper, deep neural networks are used to model speaker voice features. The GMM-UBM and the general deep neural network are used to model the speaker, and the results of the three models are compared to verify the validity of the deep neural network model. The error rate is shown in Table 1.

From the table we can see that the error rate of deep neural network is lower than GMM-UBM and general deep neural network.

| model          | Error rate |
|----------------|------------|
| GMM-UBM        | 15.23%     |
| General DNN    | 13.76%     |
| DNN            | 8.06%      |

In addition, in this paper, a dropout strategy is used to deep neural networks to suppress over-fitting of deep neural networks. The experimental results are shown in Table 2. As it can be seen from the table, after adding the dropout strategy, the network has better recognition performance.

| model          | Error rate |
|----------------|------------|
| GMM-UBM        | 15.23%     |
| General DNN    | 13.76%     |
| DNN            | 8.06%      |
| +dropout       | 6.51%      |

Theoretically, because the deep neural network is added to the RBM for pre-training, the network parameters can reach an initial optimal position during the initial training of the network. However, compared with the DNN, the general deep neural network parameters are given randomly. Figure 2 shows the iterative process of the first 1800 steps of the two neural networks. It can be seen from the figure that the initial error of the deep neural network added with the RBM is smaller than that of the general deep BP network. Moreover, since the general BP neural network is easy to fall into local minimum points, the network is oscillation. In this experiment, a dynamic learning rate adjustment algorithm is added for the deep neural network, and the degree of oscillation is significantly smaller than that of the BP network. Moreover, the overall error of the deep neural network is small, and the final convergence error is also smaller than the general deep neural network.

The above experimental results show that the deep neural network can effectively reduce the network error rate compared with the traditional speaker recognition models, and the deep neural network has stronger feature representation ability than the general deep neural network. Moreover, the dynamic learning rate adjustment, dropout and other strategies for deep neural networks can effectively optimize network performance.
4. Conclusion
In order to improve the security of the FX iDeal system, this paper uses the deep neural network with dropout strategy and dynamic learning rate algorithm to construct the user voice identification model, and compares it with the general BP neural network and GMM model. The experimental results show that the model has obvious advantages compared with the traditional models, and can exploit the speaker’s voice information better. In addition, in the FX iDeal system, a dynamic training strategy can be added to the speaker recognition model by collecting user voice data immediately. Therefore, with the data increases, the recognition performance of the network training model will continuously improve.

REFERENCES
[1] Atal B S. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4):460-475.
[2] Davis S, Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences[J]. Readings in Speech Recognition, 1980, 28(4):65-74.
[3] Reynolds D A, Quatieri T F, Dunn R B. Speaker Verification Using Adapted Gaussian Mixture Models[M]. Academic Press, Inc. 2000.
[4] Kenny P, Ouellet P, Dehak N, et al. A Study of Interspeaker Variability in Speaker Verification[J]. IEEE Transactions on Audio Speech & Language Processing, 2008, 16(5):980-988.
[5] Dehak N, Kenny P J, Dehak R, et al. Front-End Factor Analysis for Speaker Verification[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(4):788-798.
[6] Geoffrey E. Hinton,Simon Osindero,Yee-Whye Teh.A Fast Learning Algorithm for Deep Belief Nets[J].Neural computation,2006,(7):1527-1554.
[7] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6):82-97.
[8] Richardson F, Reynolds D, Dehak N. A Unified Deep Neural Network for Speaker and Language Recognition[J]. Computer Science, 2015.
[9] Lei Y, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014:1695-1699.
[10] Fischer A, Igel C. An Introduction to Restricted Boltzmann Machines[J]. 2012, 7441:14-36.
[11] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[12] Wolf J J. Efficient Acoustic Parameters for Speaker Recognition[J]. J.acoust.soc.am, 1972, 51(6).