Prospect of Voiceprint Recognition Based on Deep Learning

Shuaijie Shan¹, Jianbao Liu¹* and Yaowu Dun¹
¹College of Electrical Engineering, Naval University of Engineering, Wuhan, Hubei, 430000, China
*Corresponding author’s e-mail: jiaxy@stu.ncst.edu.cn; shanshuaijie@126.com

Abstract. As a biometric technology, voiceprint recognition is as unique as human fingerprint and pupil, so voiceprint recognition has great potential in practical application. Aiming at the feature extraction method and voiceprint recognition method, this paper first introduces the principle of voiceprint recognition, traditional MFCC, LPDA and other feature parameters and their performance; secondly, the traditional voiceprint recognition methods such as GMM and GMM-SVM are introduced, as well as their shortcomings and improvement schemes. Aiming at the shortcomings of voiceprint recognition system based on traditional algorithms in accuracy, robustness and real-time, this paper introduces the role of deep learning neural network in different stages of voiceprint recognition, and introduces the characteristics and network structure of some typical algorithms based on deep learning. Finally, according to the advantages and disadvantages of deep learning in voiceprint recognition, the development prospect and challenges of voiceprint recognition technology are analysed.

1. Introduction
As a biometric technology, voiceprint recognition can uniquely identify a person, so it is also known as speaker recognition technology[1]. Compared with other biometric technologies, it has obvious advantages of convenient data collection, simple equipment and high secrecy.

Voiceprint recognition was first proposed for the feasibility of human ear recognition mechanism and machine listening recognition[2]. Voiceprint recognition technology can be divided into speaker recognition technology and speaker confirmation technology according to different functions, and can be divided into text correlation, text restriction and text independence according to different audio content[3]. The early voiceprint recognition technology is mostly based on template matching[4], probability equation analysis, dynamic time warping[5] and other methods; feature parameter extraction is based on cepstrum, Fourier transform[6,7], MFCC and LPC[8]. In recent years, voiceprint features are mostly based on the feature vector represented by I-vector[9]; recognition is based on Gaussian mixture model and its optimization model[10,11].

Since 2014, deep learning has been gradually applied in the field of voiceprint recognition, such as DNN, CNN and LSTM models; j-vector, d-vector and x-vector feature parameters have been developed[12,13]. Then, aiming at the deep learning ability to extract highly abstract features and strong nonlinear classification ability, an end-to-end neural network structure is designed and improved to realize the integration of feature extraction and classification recognition[14,15].

This paper presents three steps of voiceprint recognition. This paper introduces the voiceprint feature parameters and their performance, the voiceprint recognition method and its principle, and the voiceprint recognition model based on deep learning. At the same time, the paper also summarizes the advantages and disadvantages of traditional recognition methods and recognition methods based on deep learning.
2. Brief analysis of voiceprint recognition

2.1. Analysis and principle of voiceprint recognition

Voiceprint refers to the spectrum of sound waves carrying voice information detected by electro-acoustics. Because of the particularity of human body structure and the complex physical process between organs, the sound is produced. Therefore, in theory, the short-term spectrum, sound source, time series, rhythm and other characteristics of each person's voice are different, that is, voiceprints are as unique and unique as fingerprints. Voiceprint recognition includes two steps: training and detection, and its technical schematic diagram is shown in Figure 1.

![Figure 1. Schematic diagram of voiceprint recognition technology](image)

2.2. Development of voiceprint recognition technology

The voiceprint features can be divided into Auditory feature that can be identified and described by human ears and Acoustic feature that can be extracted from acoustic signals by mathematical methods.

- Mel cepstrum coefficient (MFCC): MFCC is derived from the distorted frequency scale based on human auditory perception system. It is a linear transformation of logarithmic energy spectrum based on the linear Mel scale of sound frequency[16].

- Linear prediction coefficient (LPC): LPC is the characteristic parameter of analog channel structure design, which has strong robustness and high accuracy, and can approach the next sampling value through several past speech sampling values[17].

- Linear prediction cepstrum coefficient (LPCC): LPCC is the Fourier transform coefficient of the log amplitude spectrum of the linear prediction coefficient, which is obtained by calculating the spectrum envelope of LPC[18]. It can represent speech waveform and voiceprint features with limited features.

- Line spectrum frequency (LSF): LSF is the single line of line spectrum pair (LSP). The LSF model defines two resonance modes, i.e. off and on, in the connecting pipe model. Both resonance cases can be inferred based on the number of connecting tubes[19].

- Discrete wavelet transform (DWT): According to the difference of time domain and frequency domain, DWT uses different scale signal analysis. Discrete wavelet transform is an extension of wavelet transform, which can identify high-frequency signals with enhanced time resolution[20].

- Perceptual linear prediction (PLP): PLP simulates the human auditory system through a smooth short-term spectrum (which has been equalized and compressed), similar to MFCC. PLP gives the minimum resolution at high frequency and the orthogonal output similar to cepstrum analysis, and uses linear prediction to achieve spectral smoothing[21].

- Characteristics of Tandem and Bottleneck: Tandem and Bottleneck are extracted by neural network[11]. Tandem is obtained by combining posterior probability vectors corresponding to nodes of neural network output layer after dimension reduction with MFCC or PLP characteristic parameters. Bottleneck is obtained by a neural network with special structure, in which the number of nodes in one hidden layer is much less than that in other hidden layers, so this hidden layer is also called Bottleneck layer, and the output feature is Bottleneck feature.

3. Classical algorithm and analysis of voiceprint recognition

Since the application of voiceprint recognition technology, the development of voiceprint recognition technology is mainly divided into four stages, from template matching model and GMM, to the
mainstream models in voiceprint recognition field-GMM-UBM, GMM-SVM, JFA and i-Vector, and then to voiceprint recognition technology based on deep learning. They have their own advantages and disadvantages in recognition model construction and matching scores, and at the same time, they also make the channel robustness of their respective models differ[10-11].

Figure 2. Development process of voiceprint recognition technology

3.1. Template matching
Template matching is based on the common features, and automatically classifies the data to be identified into the corresponding attribute classes by using computer technology, that is, the features of the training data set and the test data set are similarly matched[22]. The most popular type is statistical pattern recognition. Template matching method can directly use the matrix of feature parameters as a template, which has low algorithm complexity and is usually suitable for text-related voiceprint recognition. Therefore, the voiceprint recognition ability under this framework is not very good.

3.2. Gaussian mixture model-general background model (GMM-UBM)
GMM is a statistical model, which can weighted average several simple normal distributions and simulate relatively complex probability distribution with fewer parameters[23]. The disadvantage of GMM is that it has too many parameters to be estimated and is very sensitive to the sound channel. In 2000, the universal background model (UBM) was introduced to model the samples outside the set. GMM was used to represent the speaker's individual characteristics, and UBM was used to represent the common characteristics. Therefore, the problem that GMM cannot be applied to open set voiceprint recognition was solved[24]. However, GMM-UBM can't compensate for the channel variability, which limits the generalization of this method.

3.3. Gaussian mixture model-support vector machine (GMM-SVM)
To solve the problem of weak anti-interference ability of GMM-UBM, WM Campbell improved GMM-UBM model by introducing SVM and its nonlinear classification ability[25]. The model constructs GSV as sample input of SVM by extracting each Gaussian component of GMM, and improves the recognition performance of GMM-UBM by using the nonlinear classification ability of SVM. In order to improve the channel anti-jamming ability, some GSV-based regularization algorithms are added to GMM-SVM, such as NAP, WCCN, etc.[26-27]. In the process of popularization, it is found that GMM-SVM is still seriously affected by channel factors in voiceprint recognition accuracy.

3.4. Joint Factor Analysis (JFA)
In order to solve the problem of too many parameters to be estimated in GMM-UBM, Canadian scientist Patrick Kenny proposed the joint factor analysis (JFA). Assuming that the speaker's space and the channel's space are independent of each other, the GSV space is divided into eigenspace and channel space[28]. The idea of JFA method is to use the subspace of GMM super vector space to model speaker difference and channel difference respectively, so as to separate channel interference. Theoretically, the absolute independence assumption seems reasonable, but there is correlation between any data; therefore,
the independence assumption limits the generalization ability of voiceprint recognition model, and it is easy to introduce errors.

3.5. i-Vector
In view of the assumption of JFA, which limits the generalization ability of voiceprint recognition model, N Dehak proposed a global difference space model, which uses full factor space to describe both speaker differences and channel differences[29]. The coordinate of speech information in the full factor space is called identity vector, and the dimension of i-vector is stable between 400 and 600. I-Vector is the best modeling framework in text-independent voiceprint recognition at present, and the subsequent improvement of voiceprint recognition technology is mostly based on the optimization of i-Vector, such as NAP, LDA, WCN and so on. At present, voiceprint recognition method based on i-Vector/PLDA model is one of the best robust models. Because i-Vector discards features including text differences, it is not as good as GMM-UBM framework in text-related recognition tasks.

4. Voiceprint recognition technology based on deep learning
Compared with traditional shallow learning, deep learning makes the classification of samples simpler. Voiceprint recognition based on deep learning greatly improves the recognition accuracy. There are two main voiceprint recognition technologies[30].

4.1. Feature Expression and Back-end Modeling Based on Deep Learning
Because the speech signal contains a variety of information and is also affected by subjective and objective factors, the appropriate voiceprint feature extraction has a great technical bottleneck. To solve this problem, the characteristics of deep neural network, such as multiple hidden layers, distributed node storage mode and deep network structure, make it have incomparable advantages in feature extraction.

4.1.1. j-Vector
Text task is to identify identity and verify speech content, but i-Vector loses a lot of relevance information in the process of dimension reduction. To this end, Nanxin Chen proposed a multi-task deep learning framework[31]. The network structure based on j-Vector is shown in Fig. 3. By extracting the average value of the last hidden layer output as a joint vector, it is represented by j-Vector.

4.1.2. d-Vector
DNN can achieve the task of speech frame classification after training. In order to further study the voiceprint recognition tasks related to small sample texts, Ehsan Variani put forward d-Vector[32]. Its working principle is: after training DNN, extract the Filterbank Energy of each frame of speech as the...
input of DNN, then extract Activations from the last hidden layer, then regularize L2, and then accumulate them to get d-Vector. In detection, the voiceprint is directly embedding into a vector, and the similarity of vectors is used to judge. The results show that this method has good performance and stronger anti-noise ability in small sample voiceprint recognition task. The schematic diagram is shown in Fig. 4.

4.1.3. x-Vector
X-Vector is an embedding feature, and scholars such as David Snyder improve the performance of DNN embedding for speaker recognition through data expansion. After training DNN, using statistics pooling layer, x-Vector can accept any length input and transform it into fixed length feature expression, that is, map the frame-level feature of speech data to the utterance-level feature[33]. As shown in fig. 5, the first four layers operate at frame level, the fifth layer is statistical information pool, and the sixth and seventh layers operate at discourse level. In addition, David Snyder introduced data enhancement consisting of additional noise and reverberation, thus expanding the amount of training data and improving robustness.

Figure 4. D-Vector model based on DNN

Figure 5. x-Vector model based on DNN
4.1.4. Backend Modeling Based on Deep Learning

The recognition method based on deep learning mainly uses deep learning to deal with voiceprint features, which makes it more distinguishable and has stronger anti-noise ability. Deep learning technology mainly uses nonlinear classification ability to compensate voiceprint features for re-channel and anti-noise, thus improving the accuracy of voiceprint recognition.

4.2. End-to-end joint optimization based on deep learning

Before the end-to-end deep neural network was put forward, the voiceprint recognition model first extracted the speaker's features from the speech data, and then used these features as input to classify or confirm the speaker. The end-to-end neural network model can realize the integration from sample input to recognition result output.

The end-to-end voiceprint recognition framework proposed by Heigold, G. combines the above three stages into a deep neural network. The input contains test statements and a small number of registration statements, and the output is simple acceptance and rejection [12]. The end-to-end architecture is shown in Fig. 6. The implementation process is to let the enrollment utterances in the lower right corner pass through the bottom DNN/LSTM to extract speech features, and then average to get the speaker recognition model; Then let the evaluation utterance in the lower right corner extract the features of the speech to be tested through DNN/LSTM; Then calculate the cosine distance between them, and finally get the recognition result.

- Logistic regression
- Score Function
- Cosine similarity
- Average
- Speaker Representation
- DNN/LSTM

Figure 6. End-to-end architecture

5. Summary of the characteristics of deep learning in voiceprint recognition

The voiceprint recognition under the traditional i-Vector framework has complex calculation and numerous parameters. The advantages of deep neural network in voiceprint recognition are as follows: Deep neural network can learn and train automatically from sample input continuously, which is simpler and more effective than i-Vector framework. The deep neural network has strong robustness and strong anti-noise ability compared with the traditional voiceprint recognition model. Deep neural network is very inclusive, which can learn all sample data and extract effective information from speech samples independently. Therefore, in the field of voiceprint recognition with complex nonlinear and multi-factor impression, deep neural network structure has high research value.

6. Development prospects and challenges

The security and industrialization of voiceprint recognition technology has always been a hot topic in this application field. The development prospects and challenges of voiceprint recognition and technology are summarized as follows.

- Deep learning and end-to-end neural network
The advantages of speaker recognition method based on deep learning are mainly embodied in speech sample training and frame-level feature extraction. How to further improve the existing deep learning method is an important aspect. And there is still much room for improvement in network structure design and data enhancement.

2) Short-term speech recognition performance
In the field of voiceprint recognition, speech signals are mostly small sample signals, and the model design for small sample voiceprint recognition is a big problem.

3) Voiceprint recognition and anti-fraud based on deep embedding learning
Facing the real and complex scenes, it is still difficult to design a robust system to solve the problems of short speech, noise, channel mismatch and large scale. Therefore, how to use deep embedding learning for voiceprint recognition and anti-fraud is also a development direction of voiceprint recognition technology.

References
[1] Lu Y.N., Shan B.Y., Guan C. (2017) Current situation and application of voiceprint recognition technology. Information system engineering, 02:13-13.
[2] Kersta, L.G. (1962). Voiceprint Identification. Nature, 196, 1253-1257.
[3] Furui, S. (1996) An overview of speaker recognition technology. Automatic Speech & Speaker Recognition, 355:31-56.
[4] O'Shaughnessy, D. (1986). Speaker recognition. IEEE ASSP Magazine, 3, 4-17.
[5] Li, W. (2017) Algorithm Research on speaker voiceprint recognition. Zhejiang University.
[6] Bogert, B., Ossanna, J.F. (1966). The heuristics of cepstrum analysis of a stationary complex echoed Gaussian signal in stationary Gaussian noise. IEEE Trans. Inf. Theory, 12, 373-380.
[7] Johnsson, S., Krawitz, R. (1992). Cooley-Tukey FFT on the Connection Machine. Parallel Comput., 18, 1201-1221.
[8] Shanthi, T. S., Lingam, C. (2014). Isolated word speech recognition system using htk. International Journal of Computer Science Engineering and Information Technology Research, 4(2), 81-86.
[9] Xin, W., Zhang, H.R. (2018) Robust i-vector speaker recognition method based on DNN processing. Computer Engineering & Applications.
[10] Li, X.K., Zheng, Y.L., et al. (2018) Research on voiceprint recognition method based on deep learning. Journal of engineering, Heilongjiang University, 009(001): 64-70.
[11] Zeng, C.Y., Ma, C.F., et al. (2020) Robust speaker recognition method based on convolutional neural network. Journal of Huazhong University of science and Technology (NATURAL SCIENCE EDITION), 48(06): 39-44.
[12] Tian, Y., Cai, M., et al. (2016) Speaker recognition system based on deep neural network and bottleneck feature. Journal of Tsinghua University (NATURAL SCIENCE EDITION), (11): 1143-1148.
[13] Lei, Y., Scheffer, N., et al. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1695-1699.
[14] Heigold, G., Moreno, I., et al. (2016). End-to-end text-dependent speaker verification. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5115-5119.
[15] Qu, F.M., Liu, J.H., et al. (2020) A Multi-fault Detection Method with Improved Triplet Loss Based on Hard Sample Mining. IEEE Transactions on Sustainable Energy, 99: 1.
[16] Xiong, B.F., Zeng, Y.C., Xie, X.J. (2016) An improved auditory feature parameter for speaker recognition. Computer applications, S1:82-85.
[17] Cai, S., Jin, X., et al. (2012) Subband energy regularized perceptual linear prediction coefficient for noise robust speech recognition. Acta acoustica Sinica, (6): 667-672.
[18] Xie, T, Zheng, X.D., et al. (2016) Analysis of seismic facies based on cepstrum coefficient of linear prediction. Acta geophysica Sinica, (11).
[19] Bai, Y.Y., Hu, X.X., Zheng S.T. (2015) Research on voiceprint recognition system based on auditory characteristics. Electronic design engineering, 23 (04): 85-87 + 91
[20] Xu, W.Y. (2011) A fast discrete wavelet transform algorithm and its application in speech signal. Computer engineering and application, 47(35): 143-146.
[21] Da, Z., Li, Q., et al. (2011) The influence of different recording systems on voiceprint detection. Journal of Nanjing University (NATURAL SCIENCE), (02): 201-207.
[22] Zhu, S.X. (2006) Research on voiceprint recognition system and pattern matching algorithm. Dalian University of technology.
[23] Lai, L.M., Hong, Q.Y. (2015) Application of voiceprint recognition in open instrument management. Journal of Huqiao University (NATURAL SCIENCE EDITION), 36(05): 517-521.
[24] Lu, X.Q., Guan, S.X. (2014) Real time voiceprint recognition based on VQ and GMM. Computer system application, 23(09): 6-12.
[25] Campbell, W., Sturim, D., & Reynolds, D. (2006). Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 13, 308-311.
[26] Solomonoff, A., Campbell, W., et al. (2005). Advances in channel compensation for SVM speaker recognition. Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., 1, 629-632.
[27] Hatch, A. O., Kajarekar, S. S., et al. (2006). Within-class covariance normalization for SVM-based speaker recognition. INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006.
[28] Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM , Montreal,(Report) CRIM-06/08-13, 14, 28-29.
[29] Dehak, N., Kenny, P., et al. (2011). Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing, 19, 788-798.
[30] Zeng, C.Y., Ma, C.F., et al. (2020) A review of speaker recognition in deep learning framework. Computer engineering and applications, 56(07): 8-16.
[31] Chen, N., Qian, Y., et al. (2015). Multi-task learning for text-dependent speaker verification. INTERSPEECH.
[32] Variani, E., Lei, X., et al. (2014). Deep neural networks for small footprint text-dependent speaker verification. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4052-4056.
[33] Snyder, D., Garcia-Romero, D., et al. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329-5333.