Overview of the development of speaker recognition

Xinyu Yuan¹, Guanyu Li¹*, Jiao Han¹, Di Wang¹, Tiankai Zhi¹

¹Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author’s e-mail: xxlgy@xbmu.edu.cn

Abstract. Speaker recognition is a technology that recognizes a person's identity through a person's voice. Speaker recognition is one of the important research topics in the field of biometrics. This article first introduces the background and research significance of speaker recognition; secondly, it briefly introduces speaker recognition technology; again introduces the method research used in speaker recognition; and finally analyzes the research difficulties and challenges of speaker recognition.

1. Introduction

Speaker recognition is also called voiceprint recognition and voice biometric recognition. Speaker recognition, as a kind of biometric authentication technology, is a technology that automatically identifies the speaker's identity based on the voice parameters that reflect the speaker's physiological and behavioral characteristics in the voice waveform[1]. Biometric identification technology is a technology that uses human physiology or behavior to identify identity. Identity authentication based on biometric identification technology is a demand for highly informatized society and economic globalization, and it is an essential technology in the government and commercial fields[2].

The biggest advantage of biometric identification technology is that the use of the human body's own characteristics as the basis for identity recognition saves users the trouble of recording text passwords or carrying documents. The current common biometric technologies include iris recognition, fingerprint recognition, palmprint recognition, gait recognition, speaker recognition. However, in addition to the speaker recognition, several other authentication methods need to be combined with professional high-cost collection equipment and require close contact. Voice is the most direct way for human communication. The development of voice-based identity authentication technology conforms to human communication habits and meets people's requirements for convenience. Speaker recognition is a technology based on voice for identity authentication, which is also a kind of biometric recognition. Compared with other identity authentication methods based on human body characteristics, voiceprint recognition has many advantages: firstly, the collection of voiceprint signals does not require high hardware equipment, and only a common microphone is needed to complete audio recording; secondly, voiceprint recognition has less personal privacy and is more humane and therefore easier to be psychologically accepted by users; thirdly, voiceprint recognition does not require people to collect information on site, making remote recognition more convenient; Furthermore, if the dynamic password is used, the voiceprint is more difficult to be copied or imitated.

2. Principles of Speaker Recognition

Both speaker recognition and speech recognition belong to the category of voice signal processing, but speaker recognition focuses on the identity information of the speaker, while speech recognition focuses
on the text information corresponding to the voice[3]. Voiceprint can be understood as the pattern of the voice frequency spectrum obtained by the time-frequency analysis technology of the waveform signal of human voice. Due to the inherent differences in the physiological structure of each person, it also causes the diversity of human speech styles, which provides us with a principle basis for automatically identifying the speaker's identity information through machines.

2.1 Speaker recognition classification
According to different application scenarios, speaker recognition can be divided into two tasks: speaker verification and speaker identification. The former refers to judging whether the current speaker is a certain identity entered in the system, which is a 1:1 confirmation question; The latter means that you don't know the identity of the current speaker, and you need to find the most similar one among the N speakers that the system has entered. It is an N:1 classification problem. According to the different recognition objects, speaker recognition can be divided into three categories: text-related, text-independent and text-prompt. The text-related speaker recognition method requires the speaker's pronunciation keywords and key sentences as training text, and the pronunciation is based on the same content during recognition. The text-independent speaker recognition method does not need to limit the speech content during training and recognition, and the recognition object is a free voice signal. The speaker recognition of text prompts, as the name suggests, the recognition object is the random generation of some specific text given by the system.

2.2 Speaker recognition process
Speaker recognition technology recognizes the identity of the speaker by analyzing the speaker characteristic information contained in the voice signal, which mainly includes two stages of training and recognition: In the training phase, according to the training speech of each speaker, feature parameters are extracted to establish a speaker model; in the recognition phase, after the speech features of the speaker to be recognized are extracted, it is matched with the established speaker model for judgment. The basic principle is shown in Figure 1.

3. Traditional methods and analysis
Gaussian Mixture Model-Universal Background Model (GMM-UBM)[5], Gaussian Mixture Model-Support Vector Machine (GMM-SVM)[6], Joint Factor Analysis (JFA)[7], i-vector[8] become the mainstream research framework in the field of speaker recognition.

3.1 GMM-UBM
The UBM model is trained on a large corpus, which includes a large number of speaker voice samples under different channels, which is equivalent to a high-mix GMM model that has nothing to do with the speaker, so it can be regarded as a fake Speaker model. The training process of the UBM model is shown in Figure 2. In the training process of the GMM-UBM model for speaker recognition, the target speaker feature is adapted to the UBM model for maximum posterior probability (MAP) adaptation to generate the corresponding target speaker GMM model.
3.2. GMM-SVM

In order to improve the anti-interference ability of the channel, Campbell introduced SVM into the modeling of GMM-UBM. Because in the GMM-UBM model, in the MAP (Maximum A Posteriori) adaptation link, only the UBM model is used to make mean adaptation to the target speaker data. Therefore, the GSV-SVM (Gaussian Super Vector) system is constructed by separately extracting the average value of each Gaussian component of GMM to form a Gaussian supervector. Relying on the powerful nonlinear classification capability of the SVM kernel function, the recognition performance is greatly improved on the basis of GMM-UBM. In addition, adding some regularization algorithms based on GSV, such as NAP (Nuisance Attribute Projection) [11], WCCN (Within Class Covariance Normalization) [12], etc., to a certain extent compensates for the influence of channel deformation on speaker modeling. However, the study found that the further improvement of its recognition rate is still seriously affected by channel factors.

3.3. JFA

In order to solve the problem of weak anti-channel interference of the GMM-UBM model, Kenny proposed JFA [13]. In JFA, the space where the speaker GSV is located is divided into eigenspace (vector features related to the speaker itself), channel space (vector features related to the channel) and residual space (vector features related to other changes). The idea of JFA is to save the features related to the speaker and remove the features related to the channel, which can well overcome the influence of the channel, and the performance of the system has been significantly improved.

Although JFA’s independent assumption of the eigen-sound space and the eigen-channel space seems reasonable, in fact the data are correlated. The absolute independent distribution assumption often provides convenience for mathematical derivation, but limits the generalization ability of the model, and each step will introduce errors during the modeling of the speaker space, channel space and residual space separately.

3.4. i-vector

i-vector is a simplified version based on JFA, which uses a Total Factor Matrix (T) to describe both speaker information and channel information, and maps the speech to a fixed and low-dimensional vector [14]. Due to the existence of channel information, it interferes with the recognition system and even seriously affects the recognition accuracy of the system. Therefore, WCCN, LDA (Linear Discriminant Analysis) [15], and PLDA (Probabilistic Linear Discriminant Analysis) [16] programs are usually used for channel compensation, noise still has a great influence on GMM characteristics.

i-vector performs well in text-independent speaker recognition, but its performance in text-related recognition is not as good as the traditional GMM-UBM framework. i-vector looks concise because it discards information such as text differences. In text-independent recognition, the content of registered speech and test speech is quite different, we need to suppress this difference; However, in text-related recognition, we need to amplify the similarity in the content of the training and recognition speech,
which causes the feature similarity of the speaker to be sparse, and the distinguishing ability decreases.

4. Deep learning methods and analysis
Deep neural network, as a modeling method in the field of pattern recognition research, not only has a strong representation ability, but also a strong classification ability. At the same time, the existence of big data provides sufficient sample support for deep learning. The task of speaker recognition is a complex classification problem. It requires that the speaker model used must have a strong representation ability, and at the same time, it must have a certain ability to distinguish the voice feature distribution of different speakers. The role of deep learning in speaker recognition can be roughly divided into three categories: feature expression, back-end modeling, and end-to-end.

The key to improving the performance of speaker recognition tasks is whether to obtain features that are rich in speaker information and contain little irrelevant information such as channel or noise. The deep learning method applied to the front end of speaker recognition can be divided into the following two types: one is the combination of DNN and traditional framework, that is, the DNN/i-vector framework; the other is to use the deep learning framework to explore a series of embedding.

4.1.DNN/i-vector
The standard i-vector framework uses UBM to align speech acoustic feature sequences and express them as high-dimensional sufficient statistics. Then map this statistic to i-vector based on factor analysis, finally, the PLDA model is used to calculate the likelihood ratio scores between different i-vectors and make a decision. In the following two subsections, we describe how to use DNN posterior probability and DNN bottleneck features in the i-vector framework.

4.1.1.DNN posterior probability
The traditional framework uses UBM to estimate the posterior probability of each Gaussian component of each frame, where each Gaussian component represents a category. These categories are obtained through unsupervised clustering, but these categories have no specific meaning and only represent space A certain area in. Using the supervised training DNN in speech recognition to replace the role of UBM, each output node of the DNN is regarded as a category. The key point is that these categories are bound triphones after clustering in speech recognition. The state has a clear correspondence with the pronunciation content, which greatly improves the performance of speaker recognition.

4.1.2.DNN Bottleneck Features
An important use of deep learning is dimensionality reduction, that is, reducing the number of hidden layer nodes in the network through nonlinear transformation and reducing the dimensionality, which can better abstract the feature details in the data. Since the number of nodes in this hidden layer is much smaller than other layers after dimensionality reduction, this layer is vividly called the BN(bottleneck) layer, and its output is called the bottleneck feature. Although the addition of the BN layer limits the flow of information within the network, it also improves the signal-to-noise ratio of the characteristic information and suppresses the redundant expression ability of the network. By suppressing the occurrence of overfitting, a better generalized model is obtained.

4.2.Embeddings
The embeddings is a representation vector that maps sentences of different lengths to a fixed dimension, so i-vector is also a kind of Embedding. However, unlike i-vector Bayesian modeling, the neural network-based embeddings method has the ability of nonlinear modeling, and embedding will be updated when the neural network is trained, and there can be many variations. The current embedding features based on DNN include d-vector[17], x-vector[18], j-vector[19], etc.

4.2.1.d-vector
In 2014, Google proposed to use a supervised training DNN to extract features, which is different from
the DNN/i-vector method, which is only used to represent the probability distribution of speech frames. Each speech frame of the speaker is input to the DNN, and the output activation value of the last hidden layer is accumulated as the representation of the specific speaker, which is the d-vector feature. In general, we use the softmax as the output layer. This method can better express the compact model of unknown speakers by removing the softmax. Unlike the traditional method, this method does not use any adaptive technology to extract the known features in the training phase, but only uses the DNN model to extract specific features in the registration and matching phase. Although simple, the d-vector is significant for the generalization of speaker recognition research in the DNN framework.

4.2.2.x-vector
x-vector is the Embeddings feature extracted by Snyder from TDNN(Time-Delay Neural Networks). Among them, the statistical pooling layer in the network structure is responsible for mapping frame-level features to segment-level or utterance-level feature, specifically calculating the mean and standard deviation of frame-level features. Since TDNN is a time-delay architecture, its network structure is characterized in that the output layer can learn long-term features, so x-vector shows stronger robustness on short speech.

4.2.3.j-vector
The extraction of i-vector relies on long speech signals, and in text-related speaker confirmation tasks, the speech is usually very short. The text-related speaker verification task should solve the problem of verifying identity and verifying voice content. j-vector (joint vector) was proposed to solve such problems. During the training of the DNN model, both the speaker and text label information are used to update the network parameters. Once the network training is completed, the two output layers are removed, and the last hidden layer is used to extract the speaker-text joint vector. In addition, the multi-task learning method avoids the over-fitting problem in DNN training and enhances the characterization ability of DNN nodes.

5.Back-end modeling based on deep learning
Deep learning acts on the back-end part of the speaker recognition system, mainly using deep learning to process traditional acoustic features, making it more distinguishable. So far, i-vector is still the best modeling framework for text-independent voiceprint recognition in most cases. The researchers' subsequent improvements are based on the optimization of i-vector. However, the previous improvement methods are based on linear methods, and the improvement of the model's distinguishing ability is limited. The deep learning method provides a non-linear way to improve the distinguishability of features by performing compensation operations on features. For example, Bhattacharya et al. proposed SCN(Speaker Classifier Network) to model speakers in the SRE2010 data set, this method uses neural networks to project i-vector into a relatively high-dimensional label space, thereby obtaining a more robust speaker representation, thereby improving the performance of the i-vector system[20].

6.Problems and challenges
Although the speaker recognition technology has made great achievements, in practical applications, short speech, background noise, channel distortion and other issues still have a great impact on speaker recognition performance. For speaker recognition technology to be widely used, the impact of various background noises on the system must be addressed. Current noise elimination algorithms cannot eliminate noise very effectively. Most anti-noise algorithms have specificity and limitations. Therefore, how to extract robust features and anti-noise models that adapt to multiple noise environments is still the focus and difficulty of future research in the field of speaker recognition technology.

Acknowledgments
This work was supported by The National Natural Science Fund (NO.61633013).
References

[1] WANG Shuzhao, QIU Tianshuang. Survey on Speaker Recognition[J]. Audio Engineering, 2007, 31(1): 51-55.

[2] Thomas Fang Zheng, Askar Rozi, Renyu Wang, etc. Biometric technology overview[J]. Information Security Research, 2016, 2(1): 12-26.

[3] BENZEGHIBAM, DE MORI R, DEROOO, et al. Automatic speech recognition and speech variability: A review[J]. Speech Communication, 2007, 49(10-11): 763-786.

[4] Chunyan Zeng, Chaofeng Ma, Zhifeng Wang, Dongliang Zhu, Nan Zhao, Juan Wang, Cong Liu. Overview of Speaker Recognition Research under the Framework of Deep Learning [J]. Computer Engineering and Applications, 2020, 56(07): 8-16.

[5] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3): 19-41.

[6] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE signal processing letters, 2006, 13(5): 308-311.

[7] Kenny P, Boulianne G, Ouellet P, et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1435-1447.

[8] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[9] Lee C H, Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters[C]// 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1993, 2: 558-561.

[10] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE signal processing letters, 2006, 13(5): 308-311.

[11] Solomonoff A, Campbell W M, Boardman I. Advances in channel compensation for SVM speaker recognition[C]// 2005 IEEE International Conference on Acoustics, Speech, and Signal (ICASSP), 2005, 1: 629-632.

[12] Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition[C]// Ninth international conference on spoken language processing, 2006.

[13] Kenny P, Boulianne G, Ouellet P, et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1435-1447.

[14] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[15] Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 155-176.

[16] Prince S J D, Elder J H. Probabilistic Linear Discriminant Analysis for Inferences About Identity[C]// 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007: 1-8.

[17] Variani E, Lei X, McDermott E, et al. Deep neural net- works for small footprint text-dependent speaker verification[C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014: 4052-4056.

[18] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]// Interspeech. 2017: 999-1003.

[19] Chen N, Qian Y, Yu K. Multi-task learning for text-dependent speaker verification[C]// Sixteenth annual conference of the international speech communication association, 2015.

[20] Bhattacharya G, Alam J, Kenn P, et al. Modelling speaker and channel variability using deep neural networks for robust speaker verification[C]// 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016: 192-198.