**Fine Tuning Speechbrain Speaker Recognition with Mandarin**

zhengkun mei(s5075580)

March 30, 2023

**Abstract**

Speaker recognition is the task of identifying persons from their voices by virtue of speech signal processing methods, and it has a broad application prospect in network communication, consumer electronics, intelligent terminals, human-computer interaction, secure payment, and other fields. There are three major branches of speaker recognition: speaker identification, speaker verification, and speaker diarization. Speaker recognition can also be divided into text-independent and text-dependent according to text information. It has been developed for decades, but it was not widely used until the rise of deep learning. At present, voiceprint recognition systems are basically based on deep learning methods, such as d-vector, x-vector, ResNet, etc. This article mainly introduces the mainstream voiceprint recognition model Emphasized Channel Attention, Propagation and Aggregation in time delay neural network Based Speaker Verification (ECAPA-TDNN), which won the first place in the International Voiceprint Recognition Competition (VoxSRC2020), and fine-tunes pretrained model which is trained on Voxceleb English dataset with zhvoice, a Mandarin dataset, to increase the performance in Mandarin speaker recognition.

1. **Introduction**
   1. **Research background**

The advancement of electronic technology has led to the development of various industries, and identity verification and identification technologies are changing rapidly in fields such as mobile payment, financial transactions, and community security. Traditional identity authentication methods are not enough to guarantee its accuracy and security. In the long-term exploration, biometric identification technology can effectively avoid the shortcomings of traditional authentication technology. The meaning of speaker speech recognition technology can be simply understood as collecting and identifying the speaker through the speaker's voice information. This is a computer discipline that intersects natural language technology and biology, and has a wide range of applications in identity verification.

In the development process of more than half a century, speaker recognition technology has shown the following advantages: low equipment cost, high crowd acceptance, good scalability, used in military defense security systems, bank security systems, and Internet security systems And it is used in public security judicial criminal investigation and identification, etc., with strong practicability and high safety. The development of artificial intelligence has led to the rapid progress of speaker recognition technology. The application of deep learning algorithms has made up for the shortcomings of traditional algorithms, greatly improved the efficiency of recognition, and expanded the application field. But at the same time, we also encountered greater challenges, such as the establishment of speech database, the loss rate of complex algorithms, model optimization and so on. Therefore, researchers are still required to further study and study to achieve rapid development of speaker recognition technology.

* 1. **The Development and Current Situation of Speaker Recognition**

The research on speaker recognition began in the 1930s [1], and the early work mainly focused on the human ear hearing discrimination experiment and exploring the possibility of listening recognition. In the 1960s, L. G. Kestar and others in the Bell Laboratory found that the spectrum of the same sound pronounced by the same person is always more similar than the spectrum of the same sound pronounced by different people through the study of spectrograms. Based on this, he used The method of visual spectrogram is used for speaker recognition, and the concept of "voiceprint" is proposed in the article [2] of the same year. Then in 1963, S. Pruzansky of Bell Labs proposed a speaker recognition method based on template matching and statistical analysis of variance [3], which attracted the attention of many scholars in the field of signal processing and raised the climax of speaker research. .

From the late 1970s to the late 1980s, the research focus of speaker recognition turned to the processing of acoustic feature parameters and new pattern matching methods. Researchers have successively proposed the Linear Predictive Coefficient (LPC) [4], the Linear Predictive Cepstrum Coefficient (LPCC) [5] (4,5 references are not necessarily accurate, read a lot of big guys The review found that the citations of these two are not the same), Mel-frequency Cepstrum Coefficient (MFCC) [6] and Perceptual Linear Predictive Coefficient (Perceptual Linear Predictive, PLP) [7] and other speaker recognition Characteristic Parameters. At the same time, dynamic time warping (Dynamic Time Warping, DTW) [8], vector quantization (Vector Quantization, VQ) [9], hidden Markov model (Hidden Markov Model, HMM) [10], artificial neural Network method (Artificial Neural Network, ANN) [11] and other technologies have also been proposed one after another, and are widely used in speaker recognition, further improving the performance of speaker recognition.

After the 1990s, especially after D. Reynolds introduced Gaussian mixture model (Gaussian mixture model, GMM) [12] in detail, GMM quickly became a model because of its simplicity, flexibility, effectiveness and good robustness. At that time, the mainstream technology in speaker recognition, which had nothing to do with text, brought speaker recognition research to a new stage. In 2000, D. Reynolds proposed the Gaussian mixture model-Universal background model GMM-UBM (Gaussian mixture model-Universal background model)[13] structure in the speaker confirmation task, making an important contribution to speaker recognition from the laboratory to the practical contribute.

In the 21st century, in the traditional GMM-UBM method, Campbell et al. found that it is very effective to use SVM for Gaussian mixture model [14]. Further, P. Kenny, N. Dehak and others successively proposed joint factor analysis (Joint factor analysis, JFA) [15] and i-vector model [16], which mapped the speaker model to a low-dimensional subspace and overcome the The limitation that the Gaussian components are independent of each other in the GMM-UBM system improves the system performance. In order to further improve the discriminative ability of the model, related discriminative training methods have also emerged (PLDA, etc., after obtaining the vector, what method do we use to compare the similarity of two vectors).

Since 2010, with the enhancement of computing power of computers, the use of deep learning methods to solve the problem of speaker recognition has become more and more important to the academic community. The deep neural network (DNN) can be used to automatically extract features, or the DNN can be used as a back-end classifier to classify the extracted features on the basis of the traditional i-vector, or an end-to-end speaker recognition network can be directly built. In 2014, Ehsan Variani[17] and others used DNN to automatically extract feature vectors from spectrograms, and named the extracted vectors d-vector. This network is very simple, that is, to send the spectrogram of each frame of a paragraph into the DNN, and then use the previous layer of the output layer as a d-vector (as for why the previous layer of the output layer is selected, this is Baidu. I remember that there is a good answer), and then average the d-vector of each frame, and the final result is the d-vecotr representing the speaker of this sentence. Then in 2015, Yu-hsin Chen[18] and others applied convolutional neural network to text-dependent speaker recognition, and achieved good results (so far, one of the two mainstreams is to use ResNet for speaker recognition ). Then, in 2017, David Snyder et al. proposed the famous x-vector [19], which was extracted on the TDNN structure, and the time pooling layer in the network was aggregated on the input speech to capture the speaker long-term characteristics. This enables the network to be trained to distinguish speakers from speech segments of different lengths (this is another mainstream method, since 2018, the baseline method of various competitions VOXSRC, NIST SRE, etc. is x-vector, and now many performances are excellent The method is to transform on this basis).

Although there have been some achievements and practical products in speaker recognition technology, there are still many problems to be solved. For example, noise will affect the recognition rate of the speaker and increase the difficulty of recognition; if the speech segment is short, the recognition effect will be significantly reduced; speech in different environments will have differences, etc. Therefore, the future of speech recognition technology still needs to invest a lot of manpower and material resources for in-depth research.

* 1. **Problem statement**

this study aims to develop a Mandarin Chinese speaker recognition model based on the ECAPA-TDNN architecture using a pre-trained English language model. The pre-trained model was trained on the Voxceleb dataset, a large-scale audio-visual dataset of human speech. The model will be fine-tuned on the zhvoice database, a Mandarin Chinese speech database, to train the Mandarin Chinese speaker recognition model.

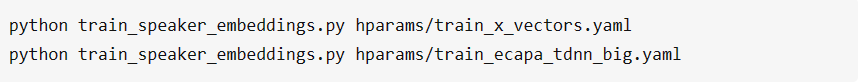
The goal of this study is to improve the accuracy of Mandarin Chinese speaker recognition by utilizing a pre-trained English language model and adapting it to Mandarin Chinese speech. This model can have a wide range of applications, from security systems to voice-controlled devices, and can provide more reliable and efficient identification for Mandarin Chinese speakers.

1. **Methodology**
   1. **Speechbrain toolkit**

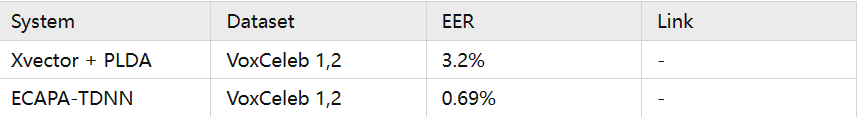
Speechbrain is a pytorch-based speech toolkit. Its ambitions are also great. It is positioned as a speech toolkit that integrates all functions. It involves various speech technologies, such as speech recognition system, speaker recognition system, speech enhancement system, multi-microphone signal processing system and so on. In this thesis, we mainly focus on the speaker recognition part.

This part of the script includes speaker recognition and speaker verification. The training and test data sets use VoxCeleb. The data is divided into two parts: Voxceleb1 and Voxceleb2. Vox1 has 1211 speakers, Vox2 has 5994 speakers, and The speakers of the two data sets do not overlap, and only one or a mixture of the two can be used during training. In addition to the original data, some data enhancement methods are also used in the script, including waveform random masking, speech rate adjustment, reverberation, noise and other operations. The final data volume is expanded to six times the original, although the training time will increase. But the effect of the model will be significantly improved.

The project provides two voiceprint models, namely X-Vector and ECAPA-TDNN (the configuration of the model structure will be described in detail later), and the training of the two models can be performed from the following code:



The project provides a pre-trained model, which is on the test set of Vox1 and Vox2, and the EER indicator is:



* 1. **Dataset**

The zhvoice corpus is composed of 8 open source data sets, processed by noise reduction and silence removal, with about 3,200 speakers, about 900 hours of audio, about 1.13 million texts, and a total of about 13 million words.

Compared with the original data, the zhvoice corpus is clearer and more natural, reducing noise interference and unnaturalness caused by incoherent speech.

The zhvoice corpus contains three aspects of information: text, speech and speaker, and can be applied to a variety of speech-related tasks.

The zhvoice corpus is cleaned and processed by Zhilang Taosha.

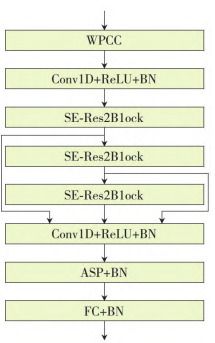
In this paper, due to the lack of time and running space, and for the reason I trained model based on the pretrained model, which need less data, I only take 100 speakers(s0002-s0102) as training data and 50 speakers(s0103-s0153) as test data.

* 1. **ECAPA-TDNN**

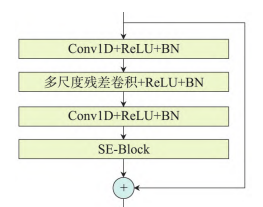
At present, the widely used speaker modeling method is the x-vector based on Time Delay Neural Network (TDNN) [20]. This method only processes the features in the last frame-level extractor, and the Level features are given equal weight. In order to improve the utilization of frame-level features and highlight the deep dynamic features with strong discrimination, the input MFCC of ECAPA⁃TDNN is replaced with WPCC containing deep dynamic features.

The structure of ECAPA⁃TDNN is shown in the figure. In the figure, Conv1D is a one-dimensional convolution operation; ReLU is a nonlinear activation function; BN is batch normalization; FC is a fully connected layer. The model uses SE⁃Res2Block to enhance the extraction ability of frame-level features, and constructs a multi-level

The remaining connection of the multi-layer feature aggregation improves the expressive ability of deep features



The SE⁃Res2Block module is shown in Figure below. This structure combines the residual structure [21] with the squeeze excitation block (Squeeze and Excitation Block, SE-Block) [22]. By adding residuals between frame-level layers connection to enhance speaker embedding features. This structure is used to model the interdependence among channels, and hierarchical residual connections are constructed to handle multi-scale features. SE⁃Block rescales the temporal context-bound frame-level features of each channel according to the global sound properties. ECAPA⁃TDNN uses multi-layer feature aggregation to fuse the final frame-level features with the first two layers of frame-level features calculated by SE⁃Res2Block to provide multi-level feature information for the statistical pooling layer. Then through the processing of the attention statistics pooling layer, the importance of each frame-level feature is given different weights, the network's attention is focused on the representative frame-level features, and the frame-level features are aggregated into segments level features. Finally, segment-level features are mapped to 512-dimensional speaker embeddings using a fully-connected layer.



* 1. **Experiment**

Voice data preparation: The zhvoice dataset is used here

Dataset size: 900 hours of Chinese voice data

Dataset Introduction:

linfo: Source data information of each dataset, including source data, introduction, etc.

ltext: The text corresponding to the speech corpus, including text, relative path, speaker, reference pinyin and other information.

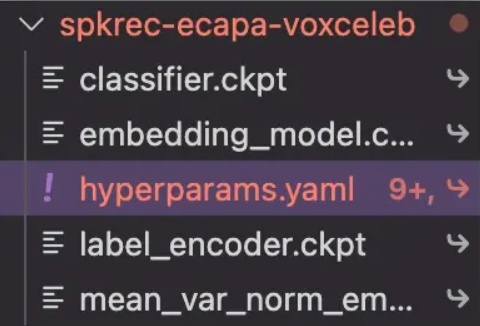
lsample: Sample speech, one audio per speaker.

lmetadata: corpus metadata, one line corresponds to one audio file, and the format audio relative path of each line\tChinese character text\n.

lzh\*: the beginning of zh is the corpus file, directory structure: the root directory contains metadata.csv and voice file directory. A speaker corresponds to a subdirectory, and the audio is in mp3 format. The data structure of metadata.csv is the same as that of metadata, which records the information of the current dataset.

Pre-training model preparation (Hugging Face)

English pre-training model



Modify the training configuration file then use colab GPU to train the model.

1. **Result**

From the following table, we can see that the EPACA-TDNN model has achieved amazing results on the zhvoice Chinese data set. In the case of zhvoice voiceprint recognition generally performing poorly (almost 50% of the EER , compared to the English database and another Chinese database ST-CMDS, which is very bad), but the zhvoice database can achieve an EER of 3.57 on the EPACA-TDNN model, which has a significant advantage over other models. And the performance is also better than most English models and Chinese models based on the ST-CMDS database.

|  |  |  |  |
| --- | --- | --- | --- |
| System | Dataset | EER | minDCF |
| I-Vector+PLDA | Voxceleb1 | 8.8 | 0.73 |
| VGG-M(softmax) | Voxceleb1 | 10.2 | 0.75 |
| VGG-M | Voxceleb1 | 7.8 | 0.71 |
| ResNet-34(1) | Voxceleb2 | 5.04 | 0.543 |
| ResNet-34(2) | Voxceleb2 | 5.11 | 0.553 |
| ResNet-34(3) | Voxceleb2 | 4.83 | 0.549 |
| ResNet-50(1) | Voxceleb2 | 4.19 | 0.449 |
| ResNet-50(2) | Voxceleb2 | 4.43 | 0.454 |
| ResNet-50(3) | Voxceleb2 | 3.95 | 0.429 |
| ResNet-34 | ST-CMDS(Mandarin) | 5.25 |  |
| EfficieNet | ST-CMDS(Mandarin) | 6.84 |  |
| GoogleNet | ST-CMDS(Mandarin) | 6.16 |  |
| ResNet-34 | Zhvoice(Mandarin) | 46.78 |  |
| EfficieNet | Zhvoice(Mandarin) | 48.26 |  |
| GoogleNet | Zhvoice(Mandarin) | 48.35 |  |
| EnsembleNet | Zhvoice(Mandarin) | 46.03 |  |
| EPACA-TDNN(with pretrain) | Zhvoice(Mandarin) | 3.57 | 0.55 |

**Conclusion**

As one of the best voiceprint recognition models, the EPACA-TDNN model can effectively improve the accuracy of voiceprint recognition. Using the pre-training model based on the English database and fine-tuning the Chinese corpus on this basis can make the model It has achieved excellent performance in the Chinese voiceprint recognition task, and its performance on the zhvoice database is far superior to other models.

**Reference**

[1]吴玺宏. 声纹识别听声辨认. 计算机世界. 2001. (8).

[2]Kersta L G. Voiceprint identification[J]. Nature, 1962, 196(4861): 1253-1257.

[3]Pruzansky, S.. “Pattern‐Matching Procedure for Automatic Talker Recognition.” Journal of the Acoustical Society of America 35 (1963): 354-358.

[4]Atal B S, Hanauer S L. Speech analysis and synthesis by linear prediction of the speech wave[J]. The journal of the acoustical society of America, 1971, 50(2B): 637-655.

[5]Atal B S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification[J]. the Journal of the Acoustical Society of America, 1974, 55(6): 1304-1312.

[6]S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust. Speech Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[7]Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. the Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752.

[8]Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, 26(1): 43-49.

[9]Burton D K, Shore J E, Buck J T. A generalization of isolated word recognition using vector quantization[C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1983, 8: 1021-1024.

[10]Rabiner L R, Juang B H. An introduction to hidden Markov models [J]. ASSP Magazine, IEEE, 1986, 3(1): 4-16.

[11]Jain A K, Mao J, Mohiuddin K M. Artificial neural networks: A tutorial [J]. Computer, 1996 (3): 31-44.

[12]D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.

[13]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Process., vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

[14]W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Process. Lett., vol.13, no. 5, pp. 308–311, 2006.

[15]Dehak N, Dumouchel P, Kenny P. Modeling prosodic features with joint factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2095-2103.

[16]Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

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

[18]Chen Y, Lopez-Moreno I, Sainath T N, et al. Locally-connected and convolutional neural networks for small footprint speaker recognition[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

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

[20] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]. Proceedings of the Interspeech 2015, 2015: 3 214- 3 218. [21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[22] HU J, SHEN L, SUN G. Squeeze and excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7 132-7 141.