Journal of Science and Engineering Research
Vol. 1, No. 1, 2022, pp. 29-46
eISSN: 2957-7187

Voiceprint Recognition based on Machine Learning Methods

Ajinkya Kunjir

Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
akujnir@lakeheadu.ca

DOI: http://dx.doi.org/10.56828/jser.2022.1.1.3
Article history: *Received (August 27, 2021); Review Result (October 30, 2021); Accepted (December 6, 2021)*

Abstract: Biometric identification technology has been widely used in today's society because of its convenience and security. As an important biometric feature, speech contains abundant information, and because of the popularity of smart devices, the collection cost of a speaker's speech is also very low. Therefore, it is of great practical value to analyze the speaker's voice. This paper mainly discusses the speaker's voiceprint recognition based on deep learning and expands the speech emotion recognition. Voiceprint recognition is divided into two tasks: speaker identification and speaker confirmation, and speech emotion recognition will be directly treated as a multi-classification problem. To take advantage of different attention mechanisms, this paper proposes a dual-path attention mechanism, which applies self-attention and convolution module attention at the same time and significantly improves the recognition effect without increasing the training time. Based on the ternary loss of predecessors, the cluster domain loss is proposed, and this paper further improves this loss for the speaker identification task and puts forward the weighted cluster domain loss, which pays more attention to the increase of the difference between classes, thus increasing the probability that critical samples are correctly classified. To solve the problem of low efficiency of cluster loss in the early stage of training, this paper also puts forward a novel loss function-critical enhancement loss, which pays extra attention to the sample pairs that are easiest and necessary to be optimized in every step of the training process. After combining cluster loss, the samples that are most difficult to optimize and easiest to optimize in every step are considered at the same time, which accelerates the training process and wins more training time for the difficult loss in cluster loss, thus further improving the final optimization effect. Aiming at the task of speaker emotion recognition, this paper proposes a lightweight neural network that combines Res Net and GRU. Compared with other methods in newer literature, this paper achieves comparable emotion classification results on the IEMOCAP data set with fewer parameters and features, in which UA reaches 67.9%, F1 score reaches 0.675, and the number of parameters is relatively reduced by 16.2%.

Keywords: Voiceprint recognition, Emotion recognition, Attention, Loss function

1. Introduction

Biometric identification technology is to identify some biological characteristics with the help of modern scientific and technological means such as computers, to achieve the purpose

of identity verification of organisms or analysis of other information. With the continuous development of modern society, biometric identification technology has been more and more used in our daily life because of its convenience and superiority. Common biometric identification technologies include fingerprint identification, iris identification, and voiceprint identification. Compared with other biological characteristics, analyzing the speaker's voice has its unique advantages. First of all, the human voice contains abundant information, and a short sentence contains various physiological characteristics of the speaker, such as gender and emotion. Besides, the uniqueness of each person's timbre also endows the voice identification function. Secondly, only one microphone is needed for voice collection, and the requirement for sound quality is relatively low. Nowadays, with the popularity of intelligent terminals, the equipment with recording functions has become very popular, and the cost of recording is relatively low. In addition, besides the speech recognition task, other applications of speech analysis only use the speaker's timbre or speech rhythm represented by the spoken sentence, and the spoken content can be flexibly changed. Therefore, for some tasks such as voiceprint recognition, the flexibility of speech brings more security and less user privacy, so it is usually more acceptable to users. It can be seen that the analysis of the speaker's voice can bring great application value. The main research content of this paper is to design a system that uses the speaker's voice to confirm or identify the identity of the speaker, and to expand the classification of voice emotion appropriately. The system can be applied in many scenarios, such as smartphone customer service scenarios, where the identity confirmation function can be used to confirm whether the speaker of the current phone is the same as the user who made the call before, or even if the user changes the mobile phone number, the user who has made the call can be found out through voiceprint comparison, and some service processes can be reduced or exempted, thus improving the efficiency of customer service personnel. By analyzing the emotional state of the current users, we can help the customer service staff to improve the current service. Speaker Recognition is also commonly called voiceprint recognition. Voiceprint recognition can be divided into two categories according to specific tasks, namely Speaker Verification (SV) and Speaker Identification (SI). Speaker verification is a 1: 1 process. In this task, the identity of the speaker to be compared is known in advance, and a test voice segment is an input. The system should compare the currently input test voice with the registered voice of the target speaker previously entered into the system to confirm whether the current test speaker identity is the target identity. Speaker identification is a 1: n process, in which a test voice segment is an input, and the system will divide this voice into the speaker who is the most similar to it among the registered identities in the system, so this task can be regarded as a multi-classification problem, and the identity of the current test voice that needs to be compared is not known in advance in this task. It should be noted that speaker identification can be divided into "closed set" and "open set". The former assumes that the speaker to be identified must be in the specified set, while the latter does not have this requirement.

In this paper, we discuss closed-set speaker identification. In addition, voiceprint recognition is divided into Text-dependent type and Text-independent type according to whether the speaker is required to speak the specified text content in the recognition process. Generally speaking, text-independent voiceprint recognition is more difficult, but its application scenarios are also more flexible, which makes it more practical. The category discussed in this paper is also the research of text-independent voiceprint recognition. Speaker emotion recognition, the so-called Speech Emotion Recognition, SER), is usually solved as a multi-classification problem, that is, a model is used to determine the emotion category to which the currently input voice belongs. The same sentence, when the speaker

expresses different emotions, may lead to differences in speech speed, tone, volume, etc., which is also the basis for the model to classify emotions. Generally speaking, the speaker emotion recognition system mainly extracts the acoustic features of the speech signal, and analyzes and summarizes the relationship between the acoustic features of the speech signal and emotion types, to complete the purpose of emotion classification of the new test speech [1]. However, it is still difficult to complete accurate emotion recognition, which is mainly due to the similar features of pronunciation in certain emotions, such as the increase of voice tone and amplitude in both happy and angry emotions. In addition, the phonetic expression boundary between some different emotions is not very obvious, which also brings difficulties to recognition.

2. Related Work

2.1. Classification of voiceprint recognition

(1) Voiceprint confirmation and voiceprint identification

According to the different sets of recognition objects, voiceprint recognition can be divided into two types: voiceprint confirmation and voiceprint recognition. Voiceprint confirmation is a one-to-one problem to determine whether a certain speech is spoken by a specific speaker. It only needs to train a model for a specific speaker, and then bring the test sample into the model and compare it with the corresponding threshold to get the test result. Its performance is not affected by the size of the training set. Voiceprint recognition is a "many-to-one" problem to determine which person in the set of people to be recognized speaks a certain speech. It is necessary to compare the test sample with all the training models and select the model with the highest matching degree as the recognition object. Therefore, the recognition rate of the voiceprint recognition system will generally decrease with the increase of models.

(2) Text-related and text-independent

According to the text relevance of training and testing speech, voiceprint recognition can be divided into two types: text-related and text-independent. Text-related voiceprint recognition requires the speaker to pronounce according to the specified training text content, and the speaker must also pronounce according to the same content of the training text when recognizing, so the recognition performance is better because the differences caused by different pronunciation contents are avoided; text-independent voiceprint recognition does not require the content of the speaker's training text, so it is more convenient and free to use and has a wider range of application scenarios. However, the technical difficulty increases and the performance is not as good as the former.

(3) Voiceprint recognition of closed set and open set

According to the data set of voiceprint recognition, it is closed set voiceprint recognition when all the speaker features contained in the test voice data exist in the training set. However, the speaker features of the test speech data in open-set voiceprint recognition do not necessarily exist in the training set. When recognizing the test samples, we should first determine whether they are in the training set, and then select the model with the highest matching degree with the test samples like closed-set voiceprint recognition, so it is more difficult to identify the open-set voiceprint, but in practical application, it is difficult to select a threshold to determine whether the test samples exist in the training set, so open-set voiceprint recognition is an urgent problem to be solved in practical application.

2.2. Research status of voiceprint recognition

The earliest use of voice to distinguish speakers occurred in the case of the death of the king of England in the 1960s. In the 1930s, people officially began to study speaker recognition, and the research work at this time mainly focused on distinguishing speakers by ear recognition. In 1932, voiceprint recognition technology was used to detect the kidnapping case of the son of an American pilot, and German intelligence agencies were also used to identify broadcasting during World War II. By the 40s, At the request of the U.S. Department of Defense, Bell Laboratories started the phonetic research and invented the spectrogram. In 1962, the laboratory used the spectrogram for artificial speaker recognition, which proved the effectiveness of the spectrogram in distinguished speakers, and put forward the concept of "voiceprint" for the first time. With the development of science and technology, Bell Laboratories engineer Pruzansky proposed a new speech recognition technology. This technology combines pattern matching with probabilistic statistical analysis of variance, setting off a new upsurge in speaker recognition, and leading researchers to shift the research focus of speaker recognition to the extraction and selection of feature parameters. In the next 20 years, researchers focused on the extraction of phonetic feature parameters. In 1963, Bogert and others successfully applied cepstrum to speaker recognition. In 1965, Tukey and Cooly first proposed a fast Fourier transform, which was widely used in the field of signal processing. In the mid-1970s, researchers found a large number of parameters that can characterize the speaker's speech features, such as linear prediction coefficient, autocorrelation coefficient, formant, cepstrum coefficient, and Mel cepstrum coefficient [1], and pointed out that cepstrum coefficient is the most effective speech feature parameter. Since the 21st century, Some scientific research institutions began to try to use prosodic information, phonemes, and other high-level information to extract information about the speaker's rhythm and accent, but this information is not stable and needs long speech, so its practicability is low. Later, researchers focused on the research of pattern matching methods. In the early 1980s, Furui combined the Mel Cepstrum coefficient with the dynamic time warping (DTW) algorithm. In 1987, Burton applied vector quantization (VQ) technology to text-related speaker recognition, with a high recognition rate. The nonparametric pattern matching method is improving, and the parametric modeling method is also breaking through. In 1989, Naik applied Hidden Markov Model (HMM) to the speaker experiment, and HMM can model the sequence well. Today HMM is still a common model in the field of speaker recognition. In the 1990s, Reynolds modeled the extracted features and proposed the Gaussian Mixture Model (GMM), which is still widely used in speaker recognition because of its easy modeling and good robustness. At the beginning of the 21st century, Vincent and others took advantage of the ability of support vector machine (SVM) to solve nonlinear problems and better represent high-dimensional feature vectors, and applied SVM to speaker recognition. Experiments proved that the effect was better than that of VQ and GMM. In 2000, Reynolds proposed Universal Background Model (UBM) for the training of GMM. This is a Gaussian mixture model trained by a large number of different speakers' voices, which is used to represent the average pronunciation features. The UBM proposal reduces the requirement of GMM for training data and enhances the robustness of the model. At present, researchers try to combine several models, and the better speaker recognition model is the combination of the support vector machine and Gaussian mixture model (SVM-GMM). The

SVM-GMM model firstly uses SVM to classify the speakers and then uses GMM to confirm the classification results [2]. After entering the 21st century, the complexity and diversity of the channel and the speaking environment have become particularly prominent, and a large number of channel compensation technologies have been used in speaker identification. Kenny proposed a joint factor analysis method in 2005 to model the channel differences. This is a method to model the differences between different speakers and different voices of the same speaker. Based on this idea, Kenny and Dehak put forward a low-dimensional vector, i-vector, which uses a low-dimensional total difference subspace to express the differences between different voice signals. Based on I-vector, Kenny further studied the application of probabilistic linear discriminant analysis from face recognition to speaker recognition and achieved good results. In 2012, Hinton's student Alex won the Image Net image recognition contest, and some scholars tried to use a neural network model for speaker recognition [3]. In recent years, artificial intelligence has entered a full-scale outbreak period. Deep learning has gradually entered the field of speaker recognition. Because of its strong autonomous learning ability and the ability to fully tap the potential information of features, deep features with high representation can be extracted by training speech samples. At present, the research focuses on the application of network structures such as deep neural network (DNN)[4-5], convolutional neural network (CNN) and recurrent neural network (RNN) in speaker recognition [6-7], and some scholars have modified the existing network to improve the robustness of speaker recognition model. Hinton and his students use a deep neural network to model voiceprint features. Compared with the traditional machine learning method, the recognition error rate is reduced by 3. 8%. Chen et al. applied the LSTM network to speaker recognition, established a multi-layer LSTM network according to the characteristics of each category, and connected the network elements between adjacent layers then, the relevance between layers is enhanced, and the classification accuracy is improved [8]. McLaren et al. combine i-vector with convolutional neural network in speaker recognition, which reduces the error rate by 26% compared with the traditional UBM+i-vector model [9]. Using a deep neural network to model speaker recognition, one research idea is to use a deep neural network to automatically extract speaker features. Another research idea is to directly use the deep neural network as a classifier, both of which have achieved success in related tasks. In practical use, Google's translation system combines with speech recognition, which can identify speakers while completing translation tasks, and it is of great practical value. Open-set speaker recognition is one of the research fields of speaker recognition, but the research on it is still rare and difficult. The bottleneck is the calculation of the threshold. Whether the selected threshold is appropriate directly affects the accuracy of model recognition. The classic thresholds include fixed threshold [10], adaptive dynamic threshold, and RS threshold. In the open set speaker recognition system, the method based on score normalization has achieved good results, and zero normalization (Z-norm) normalizes the scores of speakers outside the set of different models to the same distribution, thus eliminating the influence of differences on distribution. There are also two-level decision-making open-set speaker recognition methods, a multi-classification system combining MFCC and IMFCC parameters with a threshold algorithm. These methods have achieved certain results, but there are still many problems in practical application, such as feature parameter extraction, defects of the model algorithm, threshold calculation, and so on. Therefore, there is still much room for development in open-set speaker recognition, among which the threshold calculation is one of the research hotspots. Deep learning technology develops rapidly in speaker recognition, and related learning algorithms emerge one after another. Accelerating network training algorithm, network structure, and parameter optimization algorithm become the key factors to

improving speaker recognition effect. Traditional recognition models are difficult to dig deep personality characteristics of speakers, and the development of deep learning provides a new research direction for speaker recognition research. However, although deep learning can improve the accuracy of speaker recognition, However, if only the deep learning model is used to extract speaker features or directly used as a classifier, its performance is not ideal. Facing different speaker recognition tasks, how to choose a suitable model becomes a new problem.

2.3. Overview of machine learning

With the rapid development of computer software and hardware, statistical machine learning has been widely used in many fields. Voiceprint recognition technology is mainly based on a statistical machine learning model. Statistical machine learning is to build a probabilistic statistical model based on data. Statistical machine learning modeling is to choose a suitable model or strategy through given data and to analyze and predict unknown data with trained models, which is one of the research focuses in artificial intelligence. It has gradually become the core technology in the fields of speech, image, and so on. In statistical machine learning, the set of all features in the data set is defined as the input space, and the output of the model is defined as the output space. In the feature space, each feature is an index or a one-dimensional dimensionless vector, such as acoustic features. Sometimes, when modeling by statistical machine learning, features do not need to be mapped and transformed. However, in the research of voiceprint recognition, it is necessary to map acoustic data from input space to feature space. In a supervised classifier, it is assumed that there is a joint probability distribution function in input space, all feature vectors in this space obey this joint distribution, and all feature vectors in the training set and test set are produced according to the same distribution of probability independence. This assumption is one of the prerequisites for supervised learning about data in statistical machine learning. After defining input and output, The accuracy of model calculation is described by the loss function. The loss function commonly used in the regression model is the mean square error, while cross-entropy is commonly used as the loss function in the classification model. The loss of model training on the training set is represented by empirical risk, that is, the average loss on the training set. According to the law of large numbers, if the sample size is large enough, the difference between the calculated result of empirical risk and the expected risk will become smaller and smaller. Therefore, in the specific calculation of the model, People often use the calculation result of the empirical risk to approximate the result of the expected risk. However, due to the limited sample size in actual modeling, sometimes even small sample data set, the result is that there is a big error between empirical risk and expected risk, thus the representation ability of the model is poor. The strategy commonly used in statistical machine learning is to minimize empirical risk and structural risk. If the sample size is small or the training parameters are not set properly, The empirical risk strategy easily leads to over-fitting. To avoid over-fitting, the common method is to adopt the strategy of minimizing structural risk and add a penalty term to the model, which is the regularization term representing the complexity of the model. When training a supervised classifier, the model with small structural risk often performs better on the test set.

3. Experimental Process

The research content of this paper can be roughly divided into two parts according to specific tasks. The first part is speaker identification, that is, common voiceprint identification, and the second part is speaker emotion classification. The recognition processes of emotion recognition and speaker recognition have a lot in common, which can be represented in Figure 1.

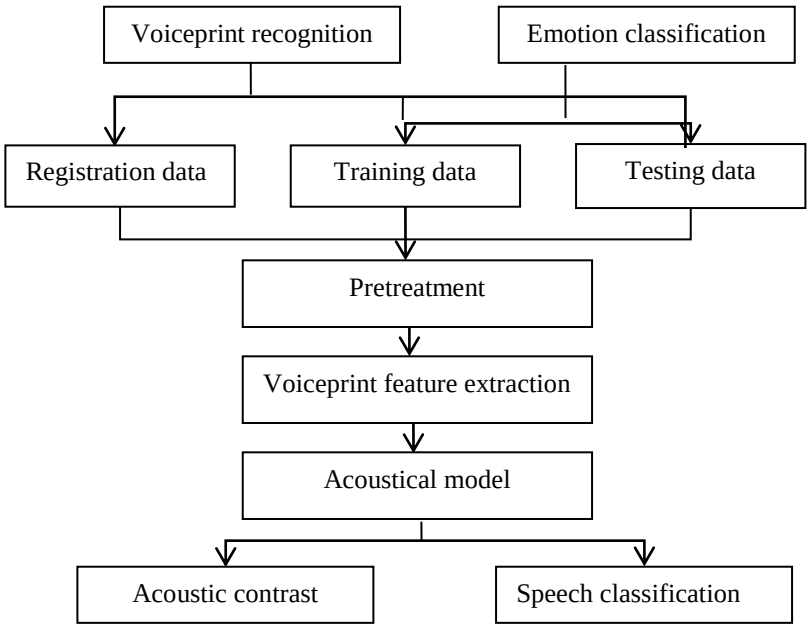

Figure 1: flow chart of speaker voiceprint recognition and emotion classification

3.1. Speech feature extraction

In the task of speech recognition, the extraction of speech features is an extremely important step. The speech signal emitted by the speaker often has great variability. If we observe it from the waveform chart, we can only know the intensity and amplitude of the speech signal, and it is difficult to obtain other useful information. Therefore, it is usually necessary to transform voice signals from the time domain to the frequency domain for analysis. The speech signal usually changes strongly, but in a very short time, such as 10-30ms, the speech signal can be roughly thought to remain unchanged, which provides the basis for the Fourier transform. At present, the most commonly used speech spectrum features are usually Melfilter-bank features or MFCC features, mainly because these two features are based on the Mel scale, imitate human auditory mechanisms, and show better performance compared with other spectrums features. [Fig.2] shows the extraction process of MFCC features and fbank features. Generally, pre-emphasis, framing, windowing, fast Fourier transform, Mel filter bank filtering, logarithm, and discrete cosine transform are required. As can be seen from Figure 2, the extraction processes of fbank features and MFCC features are very similar, and the main difference is only the last discrete cosine transform.

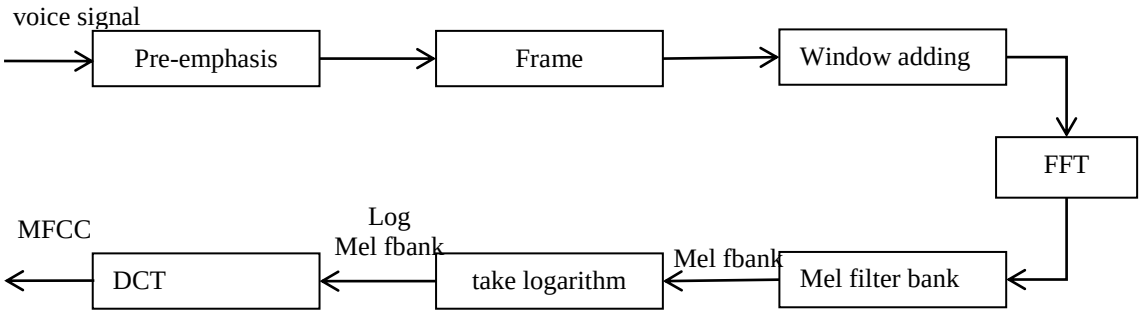

Figure 2: MFCC feature extraction flow chart

In Figure 2, the output of the Mel filter bank is actually the fbank feature, but we use the log fbank feature when we use it concretely, that is, we take the logarithm of the fbank feature again. The brighter the color in the graph, the higher the feature value, and the lower the value. It can be found that after logarithm operation, the details of the spectrum can be better highlighted, the overall texture is clearer, and it is more suitable for being input into the convolutional neural network as an image. After the logarithm operation, if we want to get MFCC features, we can do DCT transform on log fbank features. The main function of this step is to remove the coupling relationship between the features' dimensions. MFCC features are more suitable for traditional modeling methods, such as GMM. When GMM models, it is assumed that the features are uncorrelated, and the diagonal matrix is usually used instead of the full matrix. At this time, the decorrelation of DCT is very necessary. However, for the popular neural network in deep learning at present, this decorrelation transformation actually won't have a better effect but will cause the performance to decline. This is because the transformation of DCT will cause the loss of some information, and DCT is essentially a linear transformation, which is redundant in the face of DNN's powerful nonlinear fitting ability.

3.2. Introduction to data sets

3.2.1. Vox Celeb1

Voceleb1 [11] is a large English data set and one of the most commonly used standard data sets in the field of speaker recognition. The data in this dataset comes from the audio in the videos uploaded by users to Youtube. The data set contains 1,251 speakers and more than 100,000 speech segments. In addition, the gender ratio of male and female speakers in this data set is balanced and covers a wide range of contents. The average duration of each sentence in this data set is about 8s, the maximum duration is more than 2 minutes, and the

shortest duration is 4s. Generally speaking, there are many short segments. In this paper, this data set is the main data set for the speaker identification task, and the segmentation of the training set and test set for the speaker identification experiment and the speaker verification experiment are shown in Table 1 and Table 2.

Table 1: Experimental configuration of speaker identification for Vox Celeb1 data set

data set	Total number of people	Number of speech segments
Dev	1251	145,265
Test	1251	8251
Total	1251	153,516

Table 2: Speaker confirmation experiment configuration of voxceleb data set

data set	total number of people	Number of speech segments
Dev	1211	148,642
Test	40	4874
Total	1251	153,516

3.3.2. Vox Celeb2

Compared with Vox Celeb1, the Vox Celeb2 dataset [12] has greatly expanded the total amount of data, but there is no intersection between the two datasets. Vox Celeb2 dataset contains more than 1 million voice segments from more than 6,000 celebrities. The gender ratio of this data set is relatively balanced, and 61% of the speakers are male. In this paper, this data set will be used as an extended data set in the speaker verification task, and its related partition configuration is shown below. It should be noted that in this paper, we will only use the training set of the Vox Celeb2 data set, and still use the test set of the Vox Celeb1 data set during testing.

Table 3: Speaker confirmation experiment configuration of voxceleb2 data set

data set	total number of people	Number of speech segments
Dev	5994	1,092,009
Test	118	36,237
Total	6112	1,128,246

3.3.3. CNCeleb

To further verify the performance of this model on other data sets, this paper also introduces a large-scale Chinese phonetic data set recently released by Tsinghua University-CNCeleb [13]. The data set contains 1000 speakers' voices of Chinese celebrities. Because the number of individual speakers' voice files in this data set is too small (for example, some speakers only have one voice segment), which is not conducive to the experimental operation and analysis, this paper eliminated the speakers with too few voice segments in the specific experiment, leaving only 810 speakers' voice segments. The final data set is shown in [Table 4]. Compared with the Vox Celeb data set, the speech environment in CNCeleb is more

complex, such as having more background noise and more difficult to identify, so it is a challenging data set. In this paper, this data set is only used for speaker identification tasks.

Table 4: CNCeleb data set speaker identification experiment configuration

data set	total number of people	Number of speech segments
Dev	810	113,597
Test	810	12,630
Total	810	126,227

3.3.4 IEMOCAP

Interactive Emotional dyadic motion capture database, iemocap [14] data set is a data set specially used for emotional analysis, which was collected and published by Sail Laboratory of the University of Southern California. The data set contains about 12 hours of data, which includes the audio and video data of the actor's performance, the data of capturing the actor's facial expression, and the text annotation corresponding to the actor's lines. In this paper, we only discuss the emotional classification of speech, so we only use the audio data in this data set and the corresponding emotional annotation. There are nine emotional categories in this data set. In this paper, we refer to the practices in other mainstream literature [14-18], only adopt the emotional categories in five of them (happy, excited, neutral, sad, and angry), and combine happy and excited into the same category, which finally constitutes a more balanced one.

Table 5: Details of voice segments of different categories in IEMOCAP data set

Emotional category	Number of speech segments
happy	1636
angry	1103
sad	1084
neutral	1708

3.4. Training program

In this paper, the training schemes configured for each experiment are the same. During the training, we adopted the SGD gradient descent optimization method with the momentum of 0.99, and each experiment will be trained for 100 generations (because of the time problem and considering the training acceleration of CEL, the Vox Celeb2 experiment only trained for 60 generations). The decline method of learning rate is quadratic decay, and the initial learning rate is 0.005 and the final learning rate is 0.0005 in speaker recognition task training. For the emotion recognition task, the initial learning rate is 0.0003 and the final learning rate is 0.00001. The experiment is divided into speaker recognition experiment and speaker emotion recognition experiment. The speaker identification experiment is divided into speaker identification experiment and speaker verification experiment. First, we conduct the

speaker identification experiment, taking the model in reference [19] as the main comparison baseline. First, we replace the self-attention module with the dual-path attention module in this paper, and then we replace the cluster loss with the weighted cluster loss and critical enhancement loss proposed in this paper. To obtain the best weighting scheme for weighted cluster loss, this paper also discusses different weighting values. To further verify the effectiveness of this method, we also tried the above method in the speaker verification task (except for the weighted cluster domain loss). The above experiments were mainly carried out on Vox Celeb1. To explore the performance of the model on different data sets, we also carried out related experiments on Vox Celeb2 and CNCeleb data sets. In addition, to further explore the performance of the dual-path attention module proposed in this paper in other fields, we also designed a network for voice emotion recognition tasks and conducted related experiments.

3.5. Experimental results of speaker recognition

(1) The influence of different weight configurations on the WCRL effect

In this paper, the weighting coefficients in the weighted cluster domain loss are super parameters, so to determine a better weighting scheme, we experimented with different weighting values. When designing the weighted cluster domain loss, we tend to make ω2 1, so we first try to use different ω1 values when ω2 remains 1. The results of this part are shown in Table 6. It can be seen that ω1 used in the experiment can make the model get a better classification effect than the version before improvement, and the best classification effect is obtained when ω1 is 1.0004, and the accuracy rate of Top-1 is 92.0% and that of Top-5 is 97.6%.

Table 6: Speaker recognition results under different ω1 when ω2= 1 is kept on the voxceleb1 data set

Accuracy	ω1	Top-1%	Top-5%
Res-DA+WCRL	1.0006	90.9	96.9
Res-DA+WCRL	1.0005	91.2	97.3
Res-DA+WCRL	1.0004	92.0	97.6
Res-DA+WCRL	1.0003	91.2	97.2
Res-DA+CRL	1.0000	90.0	96.3

To further verify our ideas, we also tried the case where ω1< 1 or ω2 = 1. The results of this part are shown in [Table 6]. It can be seen that the performance of the model is not the best in these cases. In this case, this paper makes some guesses. First of all, ω2 should not be greater than ω1, otherwise, it will violate the original intention of the weighted cluster domain loss design in the speaker identification task in this paper, that is, pay more attention to the increase of class spacing. When ω2< ω1, the loss function can be reformulated as ω2= 1 and ω1> 1 by raising the common factor, but in this way, the value of ω1 should be redesigned. Because it can be seen from the table that the final result is sensitive to the value of ω1, we tend to keep ω2 at 1 and only adjust the value of ω1.

Table 7: Experimental results of different weighting schemes on Vox Celeb1 data set

ω1	ω2	Top-1%	Top-5%
1	1.0005	89.3	96.3
1	0.9995	89.9	96.7

1.0004	0.9995	89.4	96.5
0.9995	1	89.2	96.3
1.0004	1	92.0	97.6
1	1	90.0	96.3

(2) Comparison with other literature work

Table 8 shows the comparison between some experimental results of speaker identification in this paper and the data reported on the Vox Celeb1 data set in other literature. It can be found that the best results in this paper significantly surpass those in other literature. Compared with Res-SA+CRL in reference [19], the method proposed in this paper improves the accuracy of Top-1 by 2.9% and Top-5 by 1.8%. Previously, the method in literature [20] achieved the Top-1 accuracy of 90.8%, which was the highest recognition result on this data set. However, the best result in this paper surpassed this result, which shows the superiority of the scheme proposed in this paper.

Table 8: Comparison of results of speaker identification experiment on Vox Celeb1 data set

Accuracy	Top-1%	Top-5%
i-vector+SVM[11]	49.0	56.6
i-vector/PLDA+SVM[11]	60.8	75.6
CNN-fc-3s[11]	72.4	87.4
VGG-like CNN[11]	80.5	92.1
Res-SA[19]	85.5	93.9
VGG-like CNN+SA[20]	88.2	93.8
VGG-like CNN+CL[21]	89.5	97.0
Res Net-34+LDE[22]	89.9	95.7
Res Net-18+SA[20]	90.8	96.5
Res-SA+CRL[19]	89.1	95.8
Res-CBAM+CRL(ours)	89.5	96.0
Res-DA+CRL(ours)	90.0	96.3
Res-DA+CRL+CEL(ours)	90.2	96.7
Res-SA+CRL+CEL(ours)	90.8	97.0
Res-DA+WCRL(ours)	92.0	97.6

(3) Challenging data set testing

CNCeleb is a challenging data set. Because of its more complex voice environment, its recognition difficulty is higher than Vox Celeb1. [Table 9] shows the experimental results of the speaker identification task in this data set under different experimental configurations.

Table 9: Speaker identification results on the CNCeleb data set

Accuracy	Top-1%	Top-5%
Res-DA+CRL+CEL	84.3	92.1
Res-DA+WCRL(ω1=1.0002)	83.6	91.5
Res-DA+CRL	82.6	91.4
Res-SA+CRL	81.3	90.8

3.6. The speaker confirms the experimental results.

(1) Experimental results and comparison with other literature methods

Table 10 shows the experimental results of speaker verification on the test set of the Vox Celeb1 training set, and the comparison with other mainstream literature methods. When the self-attention in the Res-SA structure is replaced by the dual-path attention module in this paper, the equal error rate is reduced from 5.5% to 5.2%, while the critical enhancement loss plays a more important role in accelerating training, and finally, a slightly better result is obtained, which reduces the equal error rate to 5.1%. Compared with the more advanced methods such as i-vector and x-vector in mainstream methods, the scheme proposed in this paper has achieved better results.

Table 10: Effects of different models on speaker verification on Vox Celeb1 data set

model	EER
GMM-UBM[11]	15.0
i-vector-400/PLDA[11]	8.8
i-vector-2048/PLDA[23]	5.4
VGG-M[11]	7.8
x-vector(cosine)[23]	11.3
x-vector(PLDA)[23]	7.1
Res-SA + CRL[19]	5.5
Res-DA + CRL(ours)	5.2
Res-DA + CRL + CEL(ours)	5.1

In addition to EER, we also calculated the min DCF. Because there may be slight differences in the calculation parameters used in different literature, we will only compare the min DCF of the model trained by the methods mentioned in this paper. Like EER, the lower the value of min DCF, the better the model trained by this method. [Table 11] shows the min DCF values after training by different methods, which shows that the use of the dual-channel attention module and critical enhancement loss function makes the value of this index decrease.

Table 11: Min DCF values of different models on Vox Celeb1 data set

model	min DCF
Res-SA + CRL	0.621
Res-DA + CRL	0.595
Res-DA + CRL + CEL	0.588

(2) Expanded data set experiment results

Vox Celeb2 contains nearly seven times as much speech data as Vox Celeb1, which can naturally help the model to capture more speaker differences and obtain a better generalization effect. In this paper, Vox Celeb2 is directly used as the training set, and the previous verification set of Vox Celeb1 is continued. Because of the large amount of data and

long training time of Vox Celeb2, this paper only selects the best combination scheme in previous experiments, namely Resda+CRL+CEL, and considering the acceleration of cel, the experiment has only been trained for 60 generations but compared with the baseline method, it has achieved remarkable improvement effect, and the final results are compared with other literature as shown in Table 12.

Table 12: Speaker confirmation results when voxceleb2 is used as the training set

model	EER
VGG-M[11][12]	5.94
Res Net-34[12]	5.04
Res Net-50[12]	4.19
Res-SA + CRL	4.05
Res-DA + CRL + CEL	3.52

The results in Table 12 contain the results obtained from the model in the original text of Vox Celeb2 and the main baseline model Res-SA+CRL in this paper. The network parameters of ressa are only equivalent to those of resnet18, but this paper only adds a CBAM attention branch to it in the network structure. Because of the simultaneous calculation of two branches, in the experiment, we find that the extra calculation time brought by ressa is completely negligible, and the tiny calculation amount brought by the increased critical loss function will not constitute a burden on the training time. The above results can show the superiority of our model.

3.6. Speaker emotion recognition results

To verify the performance of the dual-path attention module in other applications, we experimented on speaker emotion recognition. Compared with the previous work, the structure proposed in this paper achieves better results with fewer parameters, which shows that the structure designed in this paper is concise and efficient. In addition, previous literature often uses a variety of feature combinations, such as MFCC, fundamental frequency F0, logarithmic fundamental frequency LF0, phonation probability, harmonic noise ratio, zero-crossing rate, and other features extracted in literature [18], but this paper only extracts one logarithmic F-Fbank feature, which greatly reduces the operational complexity and is closer to the requirements of real-time emotion recognition. Finally, our model achieved 67.9% UA and 0.675 F1 scores.

Table 13: Comparison of experimental results of voice emotion classification

model	Parameter quantity	UA	f1 score, f score, f measure
Xia,2017[14][18]	9.5M	60.1%	0.599
Poria,2016[15][10]	9.3M	61.3%	0.602
Poria,2017[16][18]	9.4M	57.1%	0.571
Mirsamadi,2017[12][18]	9.6M	58.8%	0.589
Runnan, 2019[18]	9.9M	67.4%	0.671
Ours	8.3M	67.9%	0.675

Table 14 shows the confusion matrix of emotion classification results. Observing the table, we can find that the model has relatively high recognition accuracy for two extreme emotions (angry and sad). However, the samples corresponding to happy and neutral emotions are more likely to be misclassified into their adjacent categories. The main reason for this is that there may not be a particularly obvious boundary between different emotions in phonetics. Except for the more extreme emotions, which are characterized by obvious features, other relatively mild emotions are prone to be misclassified.

Table 14: confusion matrix of speech emotion classification (row represents label category and column represents prediction category)

	angry	happy	neutral	sad
angry	78.2%	8.9%	9.7%	3.2%
happy	11.8%	62.7%	15.7%	9.8%
neutral	6.1%	18.8%	61.9%	13.2%
sad	3.2%	11.6%	10.5%	74.7%

4. Conclusion

In this paper, different modules are designed based on the identity and emotional recognition of the speaker's voice. According to the specific functions, the system is divided into three parts, namely, speaker recognition, speaker confirmation, and speaker emotion classification. Through experiments on several standard data sets, the effectiveness of the scheme proposed in this paper has been well proved. On speaker-text-independent voiceprint recognition and emotion classification, this paper puts forward corresponding improvement measures and shows excellent results on each data set, but there are still many improvements in the future. The main network structure adopted in this paper draws lessons from the residual network designed by predecessors, and the overall structure has been very simplified. However, if you want to run it on off-line devices, you still need to further simplify the parameters. The innovation of this paper mainly focuses on the improvement of the loss function, and the final good experimental results also show that the design of the loss function has a great disability for the optimization of the model, and further exploration of the loss function will continue in the future. Res Net and attention module used in this paper have appeared better-improved versions in other literature. If only the purpose of effect improvement is considered without considering the improvement of complexity, then the model recognition effect of this paper will be further improved after replacing the existing residual module or attention module.

Reference

[1] S. Lokesh & M. R. Devi. (2017). Speech recognition system using enhanced Mel frequency cepstral coefficient with windowing and framing method. *Cluster Computing*, 22(5), 11669-11678.

[2] I. Trabelsi, D. B. Ayed, & N. Ellouze. (2016). Comparison between GMM-SVM sequence kernel and GMM: application to speech emotion recognition. *Journal of Engineering Science and Technology*, 11(9), 1221-1233.

[3] A. Krizhevsky, I. Sutskever, & G. E. Hinton. (2017). Image net classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.

[4] M. Lee, J. Lee, & J. H. Chang. (2019). Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. *Digital Signal Processing*, 85, 1-9.

[5] Y. Le Cun, Y. Bengio, & G. Hinton. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[6] H. C. Shin, H. R. Roth, & M. Gao. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics, and transfer learning. *IEEE transactions on medical imaging*, 35(5), 1285-1298.

[7] S. Toyama, D. Saito, & N. Minematsu. (2016). Automatic estimation of extra-linguistic information in speech and its integration into recurrent neural network-based language models for speech recognition. *The Journal of the Acoustical Society of America*, 140(4), 3116-3116.

[8] R. Pradeep & K. S. Rao. (2019). Incorporation of the manner of articulation constraint in LSTM for speech recognition. *Circuits Systems and Signal Processing*, 38(8), 3482-3500.

[9] M. Mc Laren, Y. Lei, & N. Scheffer. (2014). Application of convolutional neural networks to speaker recognition in noisy conditions. *Fifteenth Annual Conference of the International Speech Communication Association*, 686-690.

[10] D. Jagadiswary & D. Saraswady. (2016). Biometric authentication using fused multimodal biometrics. *Procedia Computer Science*, 85, 109-116.

[11] A. Nagrani, J. S. Chung, & A. Zisserman. (2017). Voxceleb: A large-scale speaker identification dataset. ArXiv preprint arXiv:1706.08612.

[12] J. S. Chung, A. Nagrani, & A. Zisserman. (2018). Voxceleb2: Deep speaker recognition. ArXiv preprint arXiv:1806.05622.

[13] Y. Fan, J. Kang, & L. Li. (2020). CN-CELEB: A challenging Chinese speaker recognition dataset. *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 7604-7608.

[14] R. Xia & Y. Liu. (2015). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1), 3-14.

[15] S. Poria, I. Chaturvedi, & E. Cambria. (2016). Convolutional MKL-based multimodal emotion recognition and sentiment analysis. *IEEE 16th international conference on data mining (ICDM)*, 439-448.

[16] S. Poria, E. Cambria, & D. Hazarika. (2017). Context-dependent sentiment analysis in user-generated videos. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1, 873-883.

[17] S. Mirsamadi, E. Barsoum, & C. Zhang. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227-2231.

[18] R. Li, Z. Wu, & J. Jia. (2019). Dilated residual network with multi-head self-attention for speech emotion recognition. *ICASSP EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6675-6679.

[19] T Bian, F Chen, & L. Xu. (2019). Self-attention-based speaker recognition using cluster range loss. *Neu-recomputing*, 368, 59-68.

[20] N. N. An, N. Q. Thanh, & Y. Liu. (2019). Deep CNNs with self-attention for speaker identification. *IEEE Access*, 7, 85327-85337.

[21] S. Yadav & A. Rai. (2018). Learning discriminative features for speaker identification and verification. *Inter speech*. 2237-2241.

[22] W. Cai, J. Chen, & M. Li. (2018). Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. ArXiv preprint arXiv:1804.05160, 2018.

[23] S. Shon, H. Tang, & J. Glass. (2018). Frame level speaker embedding for text-independent speaker recognition and analysis of end-to-end mode. *IEEE Spoken Language Technology Workshop (SLT)*. 1007-1013.

This page is empty by intention.