Electronic Research Archive

ERA, 31(8): 4985–5005.
DOI: 10.3934/era.2023255
Received date: 01 March 2023
Revised date: 09 June 2023
Accepted date: 09 June 2023
Published date: 13 July 2023

http://www.aimspress.com/journal/ERA

Research article

Speech recognition of south China languages based on federated learning and mathematical construction

Weiwei Lai[1,2] and Yinglong Zheng[1,3,*]

[1] China Southern Power Grid Digital Enterprise Technology (Guangdong) Co., Ltd, Guangzhou 510000, Guangdong, China
[2] Northwestern Polytechnical University, Xi'an, Shaanxi Province, China
[3] South China University of Technology, Guangzhou, Guangdong Province, China

* **Correspondence:** Email: yinglong.zheng@voiceaitech.com.

Abstract: As speech recognition technology continues to advance in sophistication and computer processing power, more and more recognition technologies are being integrated into a variety of software platforms, enabling intelligent speech processing. We create a comprehensive processing platform for multilingual resources used in business and security fields based on speech recognition and distributed processing technology. Based on the federated learning model, this study develops speech recognition and its mathematical model for languages in South China. It also creates a speech dataset for dialects in South China, which at present includes three dialects of Mandarin and Cantonese, Chaoshan and Hakka that are widely spoken in the Guangdong region. Additionally, it uses two data enhancement techniques—audio enhancement and spectrogram enhancement—for speech signal characteristics in order to address the issue of unequal label distribution in the dataset. With a macro-average F-value of 91.54% and when compared to earlier work in the field, experimental results show that this structure is combined with hyperbolic tangent activation function and spatial domain attention to propose a dialect classification model based on hybrid domain attention.

Keywords: federated learning; South China; language speech recognition; mathematical model

1. Introduction

The goal of the research is voice recognition, which is the ability of a machine to automatically

interpret and recognize human speech using techniques like speech signal processing or pattern recognition [1]. Acoustics, pattern recognition theory, phonetics, physiology, information theory, artificial intelligence and many other fields are strongly related to speech recognition [2,3]. Speech recognition technology is increasingly important in human-computer speech communication, and its use is developing into a more competitive sector of the economy.

In the field of speech research, the most intuitive question that is usually faced is: "Who, in what language, says what?" It is important for speech research to solve this problem efficiently and with high accuracy [4]. In the key sentence above, "who" is the main object of study for voice recognition, i.e., to solve the labeling or identity determination of the key speaker; "in what language" is the main object of study for language recognition, which mainly determines the protocol criteria of this communication, i.e., What language is the main research object of language recognition, which mainly determines the agreement criteria of this communication, i.e., what language the speaker is using, i.e., Chinese or English or a local dialect, etc.; finally, "what was said" is the main research category of speech recognition [5]. Therefore, it is easy to see that language recognition, as the determiner of the agreement criteria, plays a very important qualitative role and is a key part of the speech research field.

There are several nations, ethnicities and languages in the world. The world's languages are typically divided into nine major language families according to the genealogical classification: Sino-Tibetan, Ural, Caucasian, South Island, Indo-European, Altaic, Semitic, South Asian and Dravidian [6]. Language groups, language branches and languages are further categories for the language family. The world's languages are numerous and diverse, as can be observed. According to UNESCO, there are roughly 7000 languages spoken worldwide, and nearly half of them are in danger of extinction. Dialect extinction is a serious issue because some languages have few speakers and the technology available cannot identify them. As a result, speakers slowly assimilate into other languages, learning and using some of the more popular ones as the original languages slowly disappear. As a result, research into computer-assisted language recognition and the study of language culture can, to some extent, advance human understanding of languages and language civilizations and delay the rate of those civilizations' extinction [7,8]. The process of automatically identifying the language to which a speech sample belongs using a computer or other electronic equipment is the focus of language recognition research [9,10]. As seen in Figure 1, the primary study in language recognition involves feeding the audio to be measured into a model system that has been predesigned, and then using the system's internal judgement to determine the language genus to help with other practical applications. In this study, we concentrate on the mathematics and recognition model research of speech language recognition in South China.

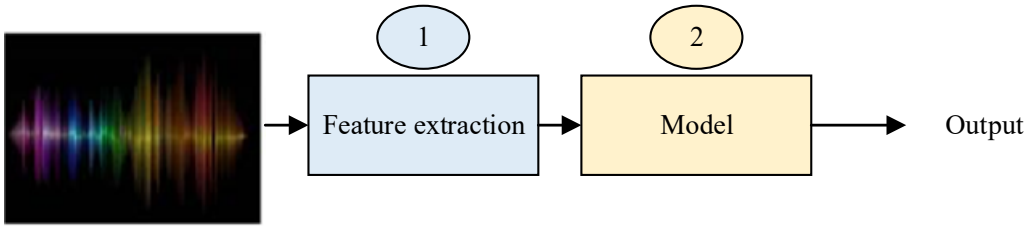

Figure 1. Speech language recognition path.

Federated learning is a new term that has emerged in recent years, evolving step by step from the

past stages of development of artificial intelligence machine learning. Initially, and currently, the most used is centralized learning, where the basic idea is "the model does not move, the data moves" [11,12], which means that the training task is performed entirely by the central server (where all the data are transferred), and this approach allows the main server to have comprehensive data and better training results. Additionally, local model parameters must be exchanged between iterations in typical distributed federated learning training methods, larger local models are needed for difficult tasks and uploading local models significantly increases the communication burden. Federated learning is frequently used on wirelessly linked devices, which dramatically increases the communication overhead. In practise, federated learning's communication burden is also made worse by network bandwidth and node density. In the classic federated learning framework, the system disables client devices that have restricted or down access during this training round, which means that the client stops optimizing updates and the server stops delivering global models to the client, which has an impact on the direct user experience. Therefore, how to reduce the communication overhead has become a key bottleneck in federated learning.

The main contribution of this paper is to design a new deep learning-based model structure and optimize it by combining the traditional speech features of South China, and the multi-task model has a higher recognition accuracy than the single-task model in the dialect language recognition problem, with an average improvement of 5%. Moreover, the model is more lightweight without involving speech information, and the deep learning features can be trained on a large scale, which provides great convenience for practicality. In summary, the speech recognition of languages in South China based on federated learning model and its mathematical model construction have important research significance.

2. Related works

A new AI learning algorithm, federated learning, was first proposed in [13]. A distributed joint averaging algorithm for efficient communication is proposed in literature [14]. The authors take the Structured updates approach to learn from a restricted space, use fewer variables for parameterization and propose Sketched updates to compress the model updates sent to the server, reducing the communication overhead.

The mapping of variable-length audio sequences into fixed-dimension vectors, which are subsequently fed into classifiers like support vector machines and probabilistic linear discriminant analysis, has been one of language recognition's more successful uses in recent years [15]. The federal model's GMM-based identification vector architecture is one of its early applications. The federal model maps speech into a whole variability space incorporating language variations and channel differences as opposed to the aforementioned approaches, which begin at the feature level, and classifies languages by transforming, deleting unnecessary information and extracting language-related information. Deep neural networks are also incorporated in the process of extracting the federation model; DNNs are used in the literature [16], replacing the role of GMM; DNNs are used in the literature [17] to extract the bottleneck features, which are then fed into the GMM-UBM model to extract the federation model; however, the aforementioned studies are still based in the GMM-UBM framework, and subsequently, with the in-depth study of deep neural networks, a framework based on front- and back-end frameworks were developed and applied to language recognition [18,19]. First DNN-based d-vector models were proposed, followed by TDNN-based x-vector models, and recently, end-to-end models have also been used in the study of language recognition, where feature extraction

and back-end classifiers are jointly trained in an end-to-end language recognition system, thus reducing errors in language classification.

More functionality and usability are the constant goals of computer technology development, and application needs are what propel this progress [20–23]. In the literature [24], DNNs were first applied to ASR systems for English in combination with HMMs, and subsequently, recurrent neural networks have all been used to mine more and richer contextual information and thus build more powerful acoustic models for English; and DNN-based ASR systems for English have reached a level comparable to human level for speech recognition tasks. End-to-end approaches have initially emerged, and attention mechanisms have been applied to ASR systems to further improve recognition performance. In addition, based on the THUYG-20 database, researchers have explored various novel ASR systems. In the literature [25], the acoustic model was improved and combined with CNN to build an acoustic model and optimize the model as a whole; in the literature [26], optimization methods and transfer learning were used to improve the recognition rate of South Chinese dialects; however, since Tsinghua University did not release a decoder for the lexical element-based language model when it released the THUYG-20 database and its baseline model, the literature [27] did not exceed the baseline model released by Tsinghua University in its attempt. In the literature [28], an attempt was made to develop a new lexeme-based language model and decoder to improve the recognition rate of the ASR system. However, the literature [29] mentions that the development of ASR for South Chinese dialects has not been as rapid as ASR for mainstream languages. South Chinese dialect is a typical dependent language, where the same stem can be connected with different affixes to form different words [30–31].

Speech recognition-related evolutions for South Chinese dialects may be somewhat hampered by language models in ASR systems running into more severe data sparsity issues, which are more likely to lead to higher out-of-word rates with a fixed vocabulary. The languages covered in this paper include many small languages, such as South China dialect, Minnan language, Hakka language, etc. Therefore, this paper constructs a low resource-based automatic speech recognition system for small languages from small languages, taking South China dialect as an example.

3. Speech recognition and mathematical

The federal model-based language recognition technique maps speech into the whole variability space, which includes linguistic variances and channel differences, and then classifies the languages by scrubbing out extraneous data and retrieving language-specific data through transformation. It was formerly the best model in the field of language recognition and has been the most effective application in the field in the last ten or so years. The federal model model is still very valuable today.

The i-vector approach is to map speech to a language- and channel-dependent total variability space and to separate it. While the predecessor of i-vector, i.e., the joint factor analysis method, an utterance can be represented as a supervector, which contains both language-related and channel-related contents, so that the supervector can be expressed as:

$$M = m + V\,y + U\,x + D \tag{1}$$

where M denotes the hypervectors from the generic background model independent of both language and channel, V and D denote the space associated with language, U denotes the space associated with channel and the vectors x, y and m denote the factors associated with language and channel.

The total variability space is a total variability matrix constructed from the feature vectors. Mapping the feature vector of speech into the total variability space, the GMM-based hypervector can be expressed as:

$$M = m + T\omega \tag{2}$$

This supervector is language and channel dependent. Where m is computed by a generic background model and is not correlated with a particular language or channel, which are often referred to as the i-vector learning model. Assuming that we have an N-frame of statements from South China $\{y_{1,}y_{2,}\ldots y_N\}$, and $UBM{:}\,\Omega$ that have been trained, and that Ω includes a total of C mixed Gaussians, then the statistics needed about extracting the i-vector can be expressed as follows:

$$N_c = \sum_{t=1}^{N} P\,(c \mid y_t, \Omega) \tag{3}$$

$$F_c = \sum_{t=1}^{N} P\,(c \mid y_t, \Omega) y_t \tag{4}$$

This process requires the calculation of the first-order statistics as follows.

$$F_c = \sum_{t=1}^{N} P\,(c \mid y_t, \Omega)(y_t - m_c) \tag{5}$$

The final i-vector mathematical model can be represented as

$$\omega = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} \tilde{F}(u) \tag{6}$$

where $1N(u)$ is a diagonal moment $CF \times CF$ of dimension $\tilde{F}(u)$, is a hypervector of dimension $CF \times 1$ and T is the total variability matrix.

Thereafter, we can construct SDC features to improve the language recognition model in South China. SDC features are one of the most widely used acoustic features in the field of language recognition at present, which are mainly extended by shift difference based on the underlying spectral parameter features MFCC or PLP. The process of SDC feature extraction for speech recognition of South China languages is shown in Figure 2.

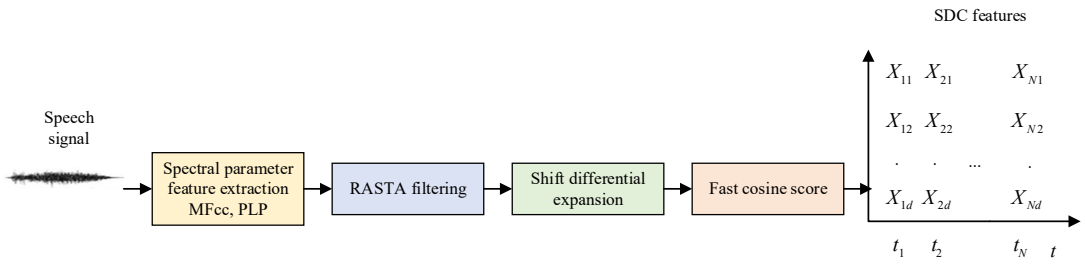

Figure 2. SDC model for language speech recognition.

First, the underlying acoustic spectral parameter features are extracted by adding windows to the speech of South China, and MFCC features or PLP features are usually extracted in the language recognition task. After extracting the spectral parameter features, RASTA filtering is adopted to suppress the influence of the non-speech spectral part of the parameter representation. In order to obtain the dynamic change features of the spectral parameters, the first- and second-order differences

(△ and △-△) of the spectral parameters are usually calculated and then spliced with the original static spectral parameter features to form the final features (for example, the 13-dimensional static MFCC features are spliced with the first- and second-order differential dynamic features to form the final 39-dimensional MFCC features). Although the use of first-order and second-order differential dynamic features has been successfully applied in the fields of speech recognition and speaker recognition, it cannot meet the requirements of language recognition tasks. This is mostly due to the statistical distribution features of the underlying acoustic units of speech, which are strongly tied to the content of speech, reflecting language information. The traditional first- and second-order differences have a short time domain extension and can reflect information that is less robust to the content, where as it is required that the underlying acoustic features can correspond to the acoustic units reflecting the speech content as much as possible. Based on this, this study proposes a time-domain shifted difference calculation method based on static spectral parameters. At this point, the resulting SDC feature is a stitching together of the static features and the k shift difference vectors $\Delta c(t,k)$ to form the final SDC feature X_t:

$$x_t = \begin{bmatrix} c(t) \\ \Delta c(t,0) \\ \Delta c(t,1) \\ \cdots \\ \Delta c(t,k-1) \end{bmatrix} \tag{7}$$

The dimension of the feature is generally $\big(N \times (K+1)\big)$. For the setting of the SDC feature parameters, the empirical parameter 7-1-3-7 is generally chosen, which is the best configuration obtained through a large number of experiments. At this time, each frame of the SDC feature calculation covers 21 frames of static parameter information of speech languages in South China, which can reflect the time duration of 210 ms (considering the window length of 10 ms when extracting static spectral parameters). It is shown that the performance of language recognition using SDC features is significantly better than that using first- and second-order differential features, and makes the acoustic feature-based approach comparable or even better than the PR-based approach.

4. Methods

This study builds a dialect classification model using the audio of South Chinese speaking languages as the research subject. Mandarin and three widely spoken dialects in the Guangdong region—Cantonese, Chaoshan and Hakka—are the focus of the research; however, there aren't any public speech datasets in the market that include these categories, so we had to gather the audio ourselves, screen it and then manually annotate it. This chapter outlines the procedure for obtaining this audio dataset, its statistical properties, and the assessment metrics utilised in this research after first introducing the background and characteristics of the Cantonese dialects that were used.

4.1. Phonological categories and distribution of features

The clear and turbid opposites of the Mandarin vowels in South China are absent, and the turbid vowels only comprise border, nasal and backward tongue fricatives. Mandarin also contains a limited

number of tones, a straightforward tonal structure, pedalized rhymes and gentle sounds. Chinese syllables can contain up to four phonemes and are often made up of vowels, rhymes and tones that form vowels and consonants separately and in a predictable way when the tones pass through the syllables. Table 1 lists all 21 vowels in order of their phonetic symbols.

Table 1. Initials of phonetic languages.

-	Labial sound				Apical				G k h		
International phonetic alphabet	$[P]$	$[P^h]$	$[m]$	$[f]$	$[t]$	$[t^h]$	$[n]$	$[I]$	$[K]$	$[K^h]$	$[X]$
Phonetic Symbols	-	-	-	-	-	-	-	-	-	-	-
Chinese Pinyin	b	p	m	f	d	t	n	l	g	k	h
-	Alveo-palatal sound				Cocky tongue sound				Lingual and dental sounds		
International phonetic alphabet	-	-	-	-	-	-	-		-	-	-
Phonetic Symbols	-	-	-	-	-	-	-		-	-	-
Chinese Pinyin	j	q	x	zh	ch	sh	r		z	c	s

The ability of native speakers of South Chinese dialects to identify whether a specific speech segment, such as the distinctive Cantonese word “pretty boy”, belongs to their own dialect suggests that the audio content already contains the necessary data for classification. Therefore, this paper classifies them by listening to the audio content. There isn’t a publicly accessible dialect dataset in China for this study that has all four categories, although it is possible to find them all in other datasets. Speech datasets have incorporated non-speech information to the audio during recording, such as device impact and background noise, which results in interference noise. There have been previous lessons that when a certain irrelevant feature in the dataset is too obviously distributed along the data category, the final learning result visualization can find that the learned classification is based on such irrelevant features and does not have generalization value. Therefore, the construction of this research dataset is unique.

For audio that is obviously noise does not help in categorization labeling, and is categorized by noise category, specifically subdivided into current noise, pure music and silence, which is left for subsequent research on noise impact or other topics; for audio mixed with multiple dialects, which is not very helpful for this topic at present, but has greater value for speaker segmentation, multi-dialect identification and other topics, it is also categorized separately and left for subsequent use; for audio containing the information required for categorization, i.e., the human ear can judge it accordingly, but the effective duration of the audio is too short, less than 1 s, i.e., the duration of the audio in which someone speaks is less than 1 s; the data set is additionally labeled, which may be helpful for the subsequent classification of short duration audio. For the audio that can be accurately categorized as a subject research category, it is additionally subdivided into the area to which the dialect belongs while categorizing it for subsequent research on the influence of region on dialect accent and other subjects. Table 2 shows the relationship between the dialects and the regions they belong to, with each category subdivided into several subcategories according to the regional differences in pronunciation. For

example, Cantonese is subdivided into five subcategories, such as Guangzhou, Siyi, Zhaoqing, Gaoyang and Wuchuan, and then the main geographical distribution of each subdivision is listed. When an audio is judged to be Cantonese, it is then divided into detailed subcategories to determine the regional accent it belongs to.

Table 2. Speech language recognition and regional correspondence.

Voice	Language classification	Region
Hakka dialect	Meizhou dialect	Meizhou, Heyuan, Huizhou, Shenzhen, Guangzhou, Dongguan, Qingyuan, Shaoguan
	Heyuan dialect	Headwater of river
	Huizhou dialect	Huizhou
	Shaoguan dialect	Shaoguan, Qingyuan
Mandarin	Cantonese accent	Guangzhou
	Guest accent	Meizhou, Shaoguan
	Chao accent	Shantou, Chaozhou, Jieyang, Shanwei
	No accent	-
Cantonese	Cantonese speech	Guangzhou, Panyu, Nanhai, Hong Kong
	Siyi dialect	Dialects in Xinhui, Enping, Kaiping, Taishan, etc.
	Zhaoqing dialect	Zhaoqing, Sihui, Luoding, Guangning, Huaiji, Fengkai, Deqing, Yunan, Yangshan, Lianxian, Lianshan and other counties and cities
	Gaoyang dialect	Maoming (Xinyi, Gaozhou), Yangjiang, Zhanjiang
	Wuchuan dialect	Wuchuan, Zhanjiang

In this study, for the dialect dataset of South China obtained through multiple rounds of screening, the audio collected in this region was classified according to Cantonese, Chaoshan and Hakka by finding native speakers of the corresponding dialects in each round of labeling.

4.2. Evaluation index calculation

In this paper, we study the topic of dialect classification and categorize the labels to be predicted for input audio in South China as one of four labels, namely, Mandarin, Cantonese, Chaoshan and Hakka. There are no audio containing two labels in the experimental dataset, so the problem can be considered as a single classification problem with multiple labels. However, the labels in the dataset are unevenly distributed, and using only accuracy or recall cannot measure the classification ability of the model for dialects, and the results are not convincing. $F_\beta Score$ widely used in the classification problem of unbalanced data sets with high reliability, the accuracy and recall are fused into a single evaluation metric by a weight ratio of $1:\beta$, calculated are shown below.

$$F_\beta = (1+\beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}} \tag{8}$$

The experiments in this paper use $F_\beta Score$ as the evaluation metric, which is the summed average of recall and accuracy, and takes values from 0 to 1, where β taken as 1 means that recall and accuracy are equally important.

In this paper, the macro-average $F1$ value, which is the largest of the two, is chosen as the main index and referred to the micro-average F1 value and accuracy rate. The macro-average first calculates the F1 value of each category separately and then arithmetic average to find the $F1$ value of the test set, which is calculated in Eq (9); the micro-average will consider all categories of prediction cases all at once and calculate the overall F1 value, which is calculated as shown below.

$$(F_1)_k = \frac{2 \times \text{precision}_k \times \text{recall}_k}{\text{precision}_k + \text{recall}_k} \tag{9}$$

$$macro(F_1) = \frac{1}{|L|}\sum_{k \in L} F_1(y, \hat{y}) = \frac{1}{|L|}\sum_{k \in L} \frac{2 \times \text{precision}_k \times \text{recall}_k}{\text{precision}_k + \text{recall}_k} \tag{10}$$

$$micro(F_1) = F_1(y, \hat{y}) = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

On this basis, the prediction for each category can be considered as one binary classification, and the four possible combinations of predicted labeling and actual labeling are shown in Table 3.

Table 3. Binary classification of speech recognition indicators.

	Actual dimensions	
Predictive annotation	TP	FP
(Expectation)	FN	TN

4.3. Model construction and operation

The model chooses support vector machines to transform the samples into another linear space using nonlinear transformation, and then determines the best classification surface in that transformation space when the sample set for South China voice recognition is nonlinear and separable. The dimensionality of the transformed space is typically higher than that of the original space because this method is utilized to linearly separate high-dimensional spaces more easily than low-dimensional ones. The computational complexity is independent of the dimensions of the space and only depends on the number of samples because support vector machines convert the original problem into a paired problem. As seen in Figure 3, this opens up the prospect of tackling high-dimensional issues.

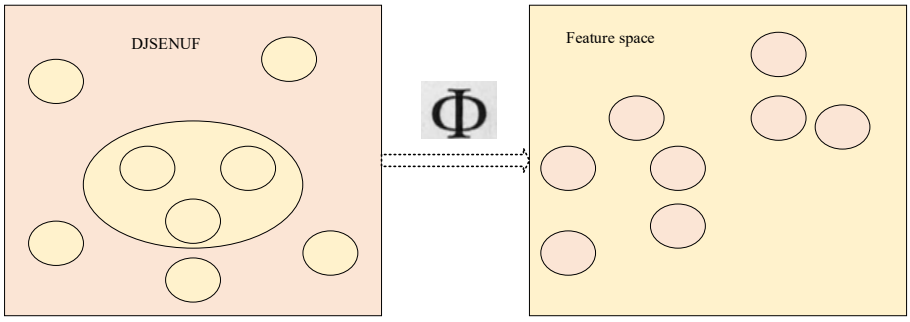

Figure 3. Input space mapping of speech recognition.

The schematic diagram of the support vector machine in the federated learning model is shown in Figure 4.

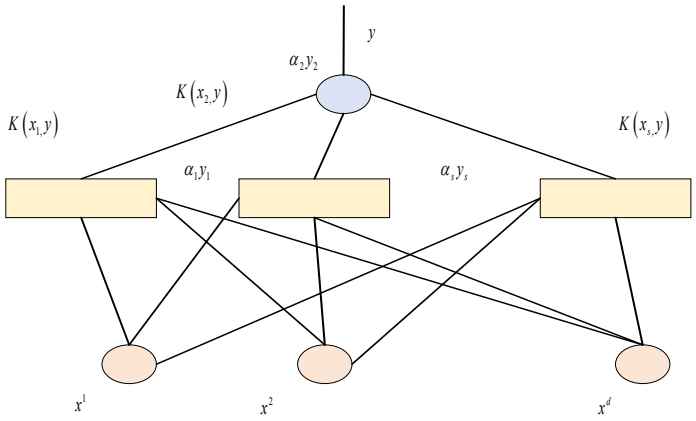

Figure 4. Mapping mode of speech recognition in federated learning.

Based on the above, the speech language recognition scheme proposed in this paper mainly consists of speech sample selection, speech pre-processing, feature parameter extraction, classifier training and testing, as shown in Figure 5.

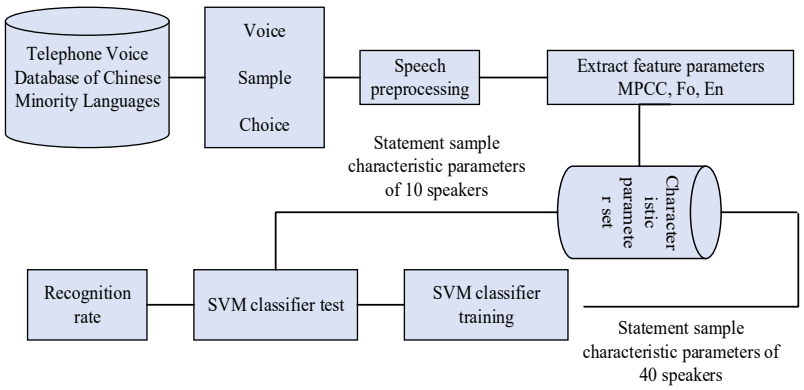

Figure 5. System block diagram of speech recognition.

The data arrangement samples for the speech language recognition model in South China are shown in Table 4. For each language, we selected 20 male speakers and 20 female speakers, each with 100 discourse samples, each with a length of approximately 60 seconds, and a total corpus size of 200 M for each language as sample data.

Table 4. Sample data arrangement of speech recognition.

Languages	Spokesman	Corpus size
Chinese	20 Men, 20 Women	40 Mb
Naxi	20 Men, 20 Women	40 Mb
Bai language	20 Men, 20 Women	40 Mb
Miao language	20 Men, 20 Women	40 Mb
Tibetan	20 Men, 20 Women	40 Mb

Intuitively, the more the number and types of feature parameters selected, the more comprehensive the speaker's linguistic features can be, but the computational cost of feature parameter extraction also increases, thus reducing the practicality of language recognition. In the experiments of the proposed ethnolinguistic language recognition scheme, different sets of feature parameters were used as the input of the classifier, i.e., the three selected feature parameters were used as the input of the classifier, and then the fusion of the three features was used as the input of the speech language recognition classifier. The experimental results are analyzed to find the best set of feature parameters with a certain correct recognition rate and minimum computational cost for ethnic languages.

4.4. Resnet-GRU based speaker recognition study

They can be identified once the mixed speech signals have been separated. Prior to speaker recognition, the speech signal must be preprocessed, which entails pre-emphasis, framing and windowing, endpoint detection and modelling using the Mel frequency spectrum's feature parameters. In this section, a speaker recognition model based on the convolutional neural network Resnet and the recurrent neural network GRU is proposed. The model uses a self-attention mechanism to assign different weights to channels with different contribution degrees, reducing the impact of redundant information and enhancing the recognition effect. A different attention mechanism module is employed than in the preceding section due to the differences in feature parameters and the volume of data required for recognition and separation.

The architecture of the speaker recognition system built in this section is shown in Figure 6:

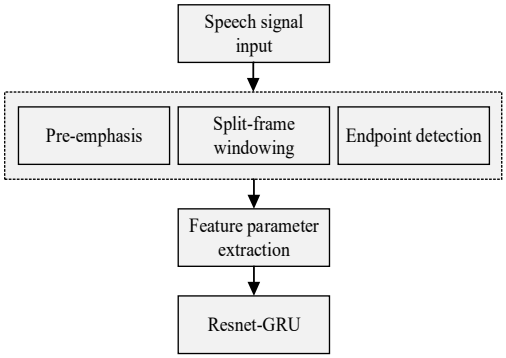

Figure 6. Functional structure of the speaker recognition system.

The functions of each part are as follows

1) Speech signal input module: the port where the speech signals from the training set and the test set enter the system, and the file input by importing the speech signals in wav format.

2) Pre-processing and feature extraction module: the input speech signal is pre-emphasized to improve the resolution of the high-frequency part and make the spectrum smoother; then, a long segment of speech signal is divided into several short segments by framing and windowing; then, the endpoint detection is used to filter out the silent segments to reduce invalid data; finally, the feature parameters are extracted and saved in Python format.

3) Model training/recognition module: Resnet-GRU model is established. The feature information of MFCC speech spectrogram is extracted by Resnet network. Although the speech signal is pre-processed, it still cannot reach the standard of noise-free and silent segments, which will lead to the degradation of recognition effect. CNN Resnet is good at extracting spatial features in the speech spectrogram, but is weak at handling temporal information in speech signals, so the recurrent neural network GRU is introduced to extract temporal features of speech signals for better speaker recognition. The traditional cross-entropy function tends to lead to huge computational overhead when training with large data volumes, so a ternary loss function is used for training, and the speech to be tested can be input to the system for recognition after the training is completed.

5. Case study

5.1. Recognition rate of speech language recognition model in South China

After pre-processing the experimental speech samples, we extracted three different types of feature parameters, including the MFCC parameter, fundamental frequency parameter and average energy parameter. We combined these feature parameter types into three different feature parameter sets as the classifier's input, and through 8-layer cross-validation, we were able to obtain the language recognition confusion matrices for the corresponding feature parameter sets of male and female speakers. By arithmetically averaging the diagonal values, which are represented in Tables 5 and 6, wherein, it represents HY Chinese, NX represents Naxi language, BY represents Bai, MY represents Miao and ZY represents Tibetan, it is possible to determine the average language recognition rates of

male and female speakers for the respective feature parameter sets.

Table 5. Male speech language recognition rate.

Languages	HY	NX	BY	MY	ZY
HY	96.37	15.00	4.17	6.67	2.31
NX	23.33	89.17	2.50	2.51	0
BY	17.50	5.00	77.50	0	2.84
MY	10.83	2.50	0	86.67	0.35
ZY	9.29	0	0.193	0.338	79.57

Table 6. Recognition rate of female and male voice languages.

Languages	HY	NX	BY	MY	ZY
HY	86.37	5.83	0	1.84	0.23
NX	16.79	97.35	0	0.45	0
BY	5.85	0	77.5	0	2.84
MY	12.78	0.83	0	88.67	1.32
ZY	2.54	0	2.91	0.43	71.53

From Tables 5 and 6, the recognition rate of fundamental frequency is the highest in the single feature experiment, and the recognition rate of dialect languages in South China is also very good, while the recognition rate of some dialects in remote areas is slightly lower, but also reaches more than 70%. The basic frequency parameter characteristic is more likely to be recognized by male and female speakers than other aspects. As a result, language recognition systems frequently use base-frequency features. The fusion of the three chosen features performed better and had a higher identification rate than utilizing a single feature in the experiment, hence it is worthwhile to further investigate this method. However, combining more features is not recommended because doing so will result in redundant feature data and worse language recognition rates, as illustrated in Figure 7.

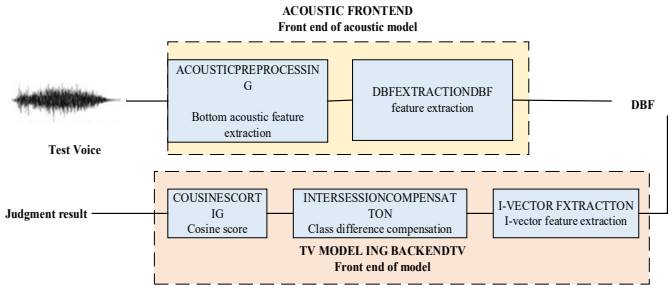

Figure 7. Flow chart of DBF-TV system with relevant features of voice.

After the recognition rate of this model is trained, the backend classification is also trained from the same training set and the i-vectors of the training set need to be extracted before training the backend. The backends used in this paper include cosine distance, logistic regression (LR), PLDA and SVM. After the backend training is completed, the i-vectors of the training set are normalized

according to the language, representative features of each language are extracted and the normalized features of each language in the training set are compared with the i-vectors of the test and validation sets, and the classification is scored. After the model testing is completed, the whole experimental optimization process is shown in Figure 8.

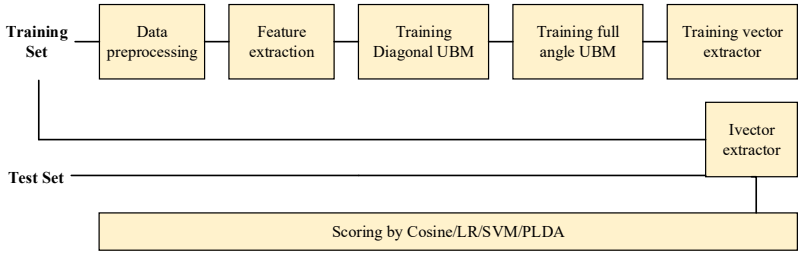

Figure 8. Optimization model of speech recognition training.

The performance of the model on the speech language recognition database in South China is shown in Table 7.

Table 7. Speech recognition results.

Model	Test set		Validation set	
	C_{avg}	EER%	C_{avg}	EER%
I-vector cosine	0.0692	7.83	0.05	6.19
I-vector LR	0.0542	5.53	0.045	4.55
I-vector RBF SVM	0.0559	5.47	0.034	3.23
I-vector PLDA	0.0505	7.51	0.0339	5.17

According to the test results, it can be seen that due to the different selection of classifiers, the results have great differences although they are the same i-vector, among which SVM as the back-end classifier has the best recognition effect. Specifically, for each language, different backends, although there is a big difference in the results, such as i-vector+SVM has an EER of 5.47% on the test set; i-vector+cosine has an EER of 7.83% on the test set, but the classification ability for each language is similar, only the recognition rate in the current language has improved performance varies, as shown in Figure 9 in the confusion matrix III is shown.

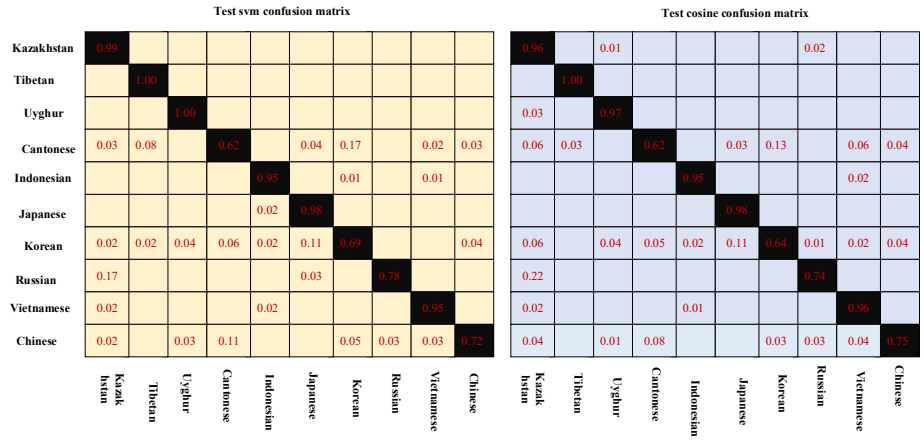

Figure 9. The confusion matrix of speech recognition test.

According to the experimental results, it can be found that the recognition accuracy of the validation set outperforms the test set on different backends in the Oriental language database, and the difference between the two lies in whether the speaker is the same as the training set, and the speaker mismatch causes a significant performance degradation of the system.

5.2. Optimization efficiency of speech language recognition models in South China

The data is processed before the model is trained, including returning pairs (both positive and negative pairs) and target values. For the training set: a positive pair or a negative pair is created randomly for each sample. If target = 1, then select training data 2 with the same label as training data 1 to form a positive pair; if target = 0, then select training data 2 with labels other than those belonging to training data 1 to form a negative pair. For the validation set: create fixed pairs for testing. Iterate through the training data of the validation set (N) in steps of 2 to generate N/2 positive pairs and N/2 negative pairs, respectively. If positive pairs are generated, target = 1; if negative pairs are generated, target = 0. For the test set: create fixed pairs for testing and iterate through the training data (N) of the test set in order, in steps of 2, to generate N/2 positive pairs and N/2 negative pairs, respectively. If positive pairs are generated, target = 1; if negative pairs are generated, target = 0. The recognition results are shown in Table 8.

Table 8. Twin network recognition results of speech languages.

Database	Model	Validation set	Test set
-	-	EER%	EER%
Oriental languages	X-vector DNN	5.02	6.67
-	X-vector Siamese	4.79	5.67
Broadcast voice	X-vector DNN	-	13.44
-	x-vector Siamese	-	13.0

Following the method described in Section 3, this section tests the recognition accuracy of

different models trained on the test set when the set of Fbank feature vectors is used as input. The specific models include LDnn built by DNN, LCnn built by CNN and LClstm, LCgru based on CNN and RNN as mentioned above. The recognition results of each model are shown in Table 9. We can see that the recognition accuracy of LClstm is the best, but at the same time this network is also the most complex network structure.

Table 9. The accuracy of different optimization model structures.

Network structure	Accuracy
Baseline system	77.8%
LSTM	83.9%
GRU	84.3%
CNN	85.2%
CNN+LSTM	89.2%

In addition, in the task of multiple language recognition, the model can define multiple tasks, i.e., recognizing each dialect language as a task, and each task corresponds to a loss function, and jointly training multiple loss functions, so that the multi-task model learns the implicit features that are easily ignored among each other, including intonation, pitch, curl and other features. MTLNet, a multilingual task dialect language recognition model for South China, is built, and the optimized model structure is shown in Figure 10.

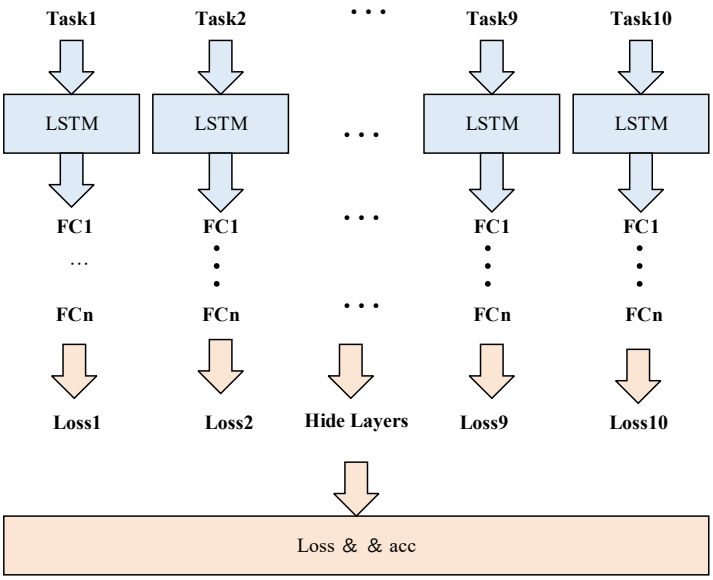

Figure 10. Optimization model of speech multitask recognition.

After optimization of the model, we apply the same data input and feature extraction strategy to the speech language recognition task and auxiliary task in South China, and we can get Figure 11.

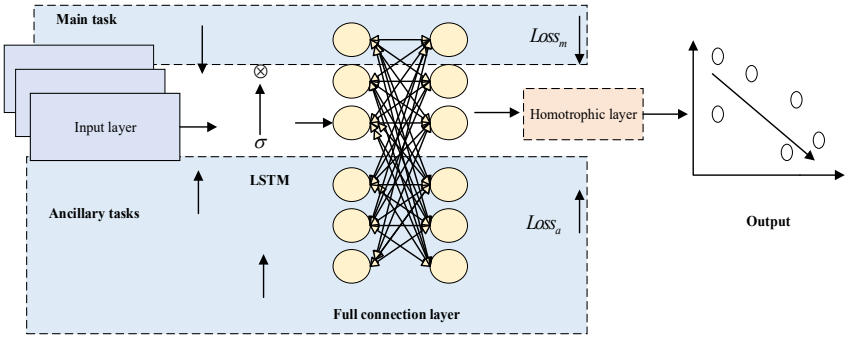

Figure 11. Speech assistant task recognition model.

After model optimization, the multilingual task model uses a single language recognition task from South China for each subtask of multilingual recognition, and the multitask neural network is jointly trained. The network was enhanced by linearly increasing the number of tasks, and after each increment, it was retrained, tested and its performance was statistically assessed. The experimental results in Figures 12 and 13 illustrate a comparison and analysis between the single-task dialect language recognition model and the multilingual task dialect language recognition model.

From Figure 13, it can be seen that the increase of data dimensions in single-task network and the increase of task dimensions in multi-task network can both reduce the loss and improve the model recognition accuracy. In the single-task model, the improvement of accuracy mainly comes from the increase of input data dimensions to make the training samples richer [32].

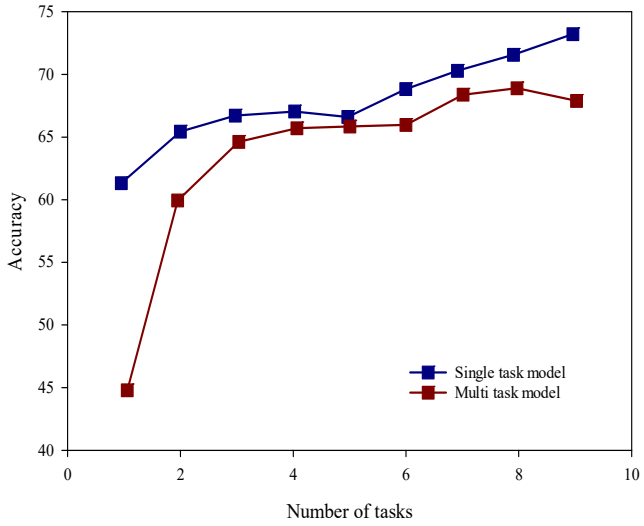

Figure 12. Comparison results of speech recognition.

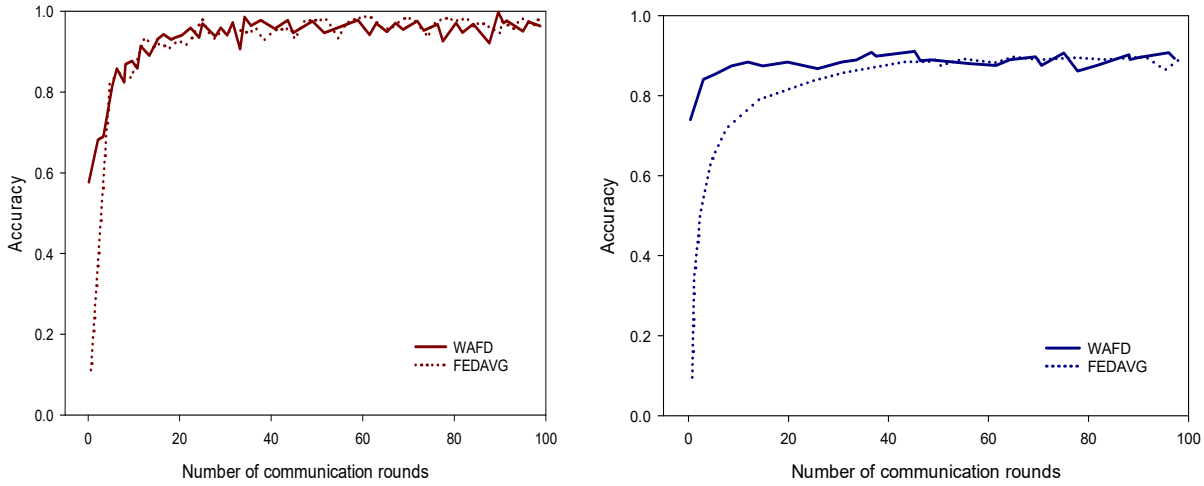

Figure 13. Comparison results of speech recognition communication rounds.

6. Conclusions

This paper integrates established techniques for speech language identification in South China, creates fresh deep learning model architectures and enhances their usefulness. In the meantime, strategies that are appropriate for the investigation of dialect recognition with high similarity and simple confusion are put forth. The experimental findings demonstrate that combining attention mechanism, data enhancement and tuning reference for model optimization ultimately leads to a good fitting of the language recognition neural network, with the recognition accuracy reaching about 98% for two languages and 90% for five languages. The multi-task model also performs better than the single-task model in the dialect language recognition problem, with an average improvement of 5%.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are sincerely grateful to the Academic Editor for the careful reading, updated information about references, detailed comments and valuable suggestions that make the manuscript improved substantially after revision.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Process Mag.*, **29** (2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597
2. E. T. Affonso, R. D. Nunes, R. L. Rosa, G. F. Pivaro, D. Z. Rodriguez, Speech quality assessment in wireless voip communication using deep belief network, *IEEE Access*, **6** (2018), 77022–77032. https://doi.org/10.1109/ACCESS.2018.2871072
3. B. Alekhya, R. Sasikumar, An ensemble approach for healthcare application and diagnosis using natural language processing, *Cognit. Neurodyn.*, **16** (2022), 1203–1220. https://doi.org/10.1007/s11571-021-09758-y
4. J. H. Hansen, T. Hasan, Speaker recognition by machines and humans: A tutorial review, *IEEE Signal Process Mag.*, **32** (2015), 74–99. https://doi.org/10.1109/MSP.2015.2462851
5. D. Li, Z. Luo, B. Cao, Blockchain-based federated learning methodologies in smart environments, *Cluster Comput.*, **25** (2022), 2585–2599. https://doi.org/10.1007/s10586-021-03424-y
6. T. Samad, J. S. Bay, D. Godbole, Network-centric systems for military operations in urban terrain: The role of UAVs, *Proc. IEEE*, **95** (2007), 92–107. https://doi.org/10.1109/JPROC.2006.887327
7. Y. Bai, Y. Zhao, Y. Shao, X. Zhang, X. Yuan, Deep learning in different remote sensing image categories and applications: status and prospects, *Int. J. Remote Sens.*, **43** (2022), 1800–1847. https://doi.org/10.1080/01431161.2022.2048319
8. J. C. Zhou, J. M. Sun, W. S. Zhang, Z. F. Lin, Multi-view underwater image enhancement method via embedded fusion mechanism, *Eng. Appl. Artif. Intell.*, **121** (2023), 105946. https://doi.org/10.1016/j.engappai.2023.105946
9. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: A systematic review, *IEEE Access*, **7** (2019), 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880
10. M. Kubanek, J. Bobulski, J. Kulawik, A method of speech coding for speech recognition using a convolutional neural network, *Symmetry*, **11** (2019), 1185. https://doi.org/10.3390/sym11091185
11. G. E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Trans. Audio Speech Lang. Process.*, **20** (2011), 30–42. https://doi.org/10.1109/TASL.2011.2134090
12. Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concepts and applications, *ACM Trans. Intell. Syst. Technol.*, **10** (2019), 1–19. https://doi.org/10.1145/3298981
13. Y. Liu, Y. Kang, C. Xing, T. Chen, Q. Yang, A secure federated transfer learning framework, *IEEE Intell. Syst.*, **35** (2020), 70–82. https://doi.org/10.1109/MIS.2020.2988525
14. C. Nadiger, A. Kumar, S. Abdelhak, Federated reinforcement learning for fast personalization, in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, **9** (2019), 123–127. https://doi.org/10.1109/AIKE.2019.00031
15. K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, et al., Secureboost: A lossless federated learning framework, *IEEE Intell. Syst.*, **36** (2021), 87–98. https://doi.org/10.1109/MIS.2021.3082561
16. S. Zhang, L. Yao, A. Sun, A. Sun, Deep learning-based recommender systems: a survey and new perspectives, *ACM Comput. Surv.*, **52** (2019), 1–38. https://doi.org/10.1145/3285029

17. S. S. Khanal, P. W. C. Prasad, A. Alsadoon, A. Maag, A systematic review: machine learning based recommendation systems for e-learning, *Educ. Inf. Technol.*, **25** (2020), 2635–2664. https://doi.org/10.1007/s10639-019-10063-9
18. Z. Batmaz, A. Yurekli, A. Bilge, C. Kaleli, A review on deep learning for recommender systems: challenges and remedies, *Artif. Intell. Rev.*, **52** (2019*)*, 1–37. https://doi.org/10.1007/s10462-018-9654-y
19. P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, P. Swietojanski, Adaptation algorithms for neural network-based speech recognition: An overview, *IEEE Open J. Signal Process.*, **2** (2021), 33–66. https://doi.org/10.1109/OJSP.2020.3045349
20. J. C. Zhou, L. Pang, D. Zhang, W. S. Zhang, Underwater image enhancement method via multi-interval subhistogram perspective equalization, *IEEE J. Oceanic Eng.*, **48**2023), 474–488. https://doi.org/10.1109/JOE.2022.3223733
21. M. T. Patrick, K. Raja, K. Miller, J. Sotzen, J. E. Gudjonsson, J. T. Elder, et al., Drug repurposing prediction for immune-mediated cutaneous diseases using a word-embedding–based machine learning approach, *J. Invest. Dermatol.*, **139** (2019), 683–691. https://doi.org/10.1016/j.jid.2018.09.018
22. L. Li, Y. Wang, K. Y. Lin, Preventive maintenance scheduling optimization based on opportunistic production-maintenance synchronization, *J. Intell. Manuf.*, **32** (2021), 545–558. https://doi.org/10.1007/s10845-020-01588-9
23. S. Lloyd, C. Weedbrook, Quantum generative adversarial learning, *Phys. Rev. Lett.*, **121** (2018), 040502. https://doi.org/10.1103/PhysRevLett.121.040502
24. H. Kim, J. Park, M. Bennis, S. L. Kim, Blockchained on-device federated learning, *IEEE Commun. Lett.*, **24** (2019), 1279–1283. https://doi.org/10.1109/LCOMM.2019.2921755
25. T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, W. Shi, Federated learning of predictive models from federated electronic health records, *Int. J. Med. Inf.*, **112** (2018), 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
26. P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, et al., Advances and open problems in federated learning, *Found. Trends Mach. Learn.*, **14** (2021), 1–210. https://doi.org/10.1561/2200000083
27. C. Zhang, M. Li, D. Wu, Federated multidomain learning with graph ensemble autoencoder GMM for emotion recognition, *IEEE Trans. Intell. Transp. Syst.*, **24 (**2023**),** 7631–7641. https://doi.org/10.1109/TITS.2022.3203800
28. J. Men, G. Xu, Z. Han, Z. Sun, X. Zhou, W. Lian, et al., Finding sands in the eyes: vulnerabilities discovery in IoT with EUFuzzer on human machine interface, *IEEE Access*, **7** (2019), 103751–103759. https://doi.org/10.1109/ACCESS.2019.2931061
29. S. Truex, L. Liu, M. E. Gursoy, L. Yu, W. Wei, Demystifying membership inference attacks in machine learning as a service, *IEEE Trans. Serv. Comput.*, **14** (2019), 2073–2089. https://doi.org/10.1109/TSC.2019.2897554
30. M. Shen, H. Wang, B. Zhang, L. Zhu, K. Xu, Q. Li, et al., Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing, *IEEE Internet Things J.*, **8** (2020), 2265–2275. https://doi.org/ 10.1109/JIOT.2020.3028110
31. S. W. Graham, R. G. Olmstead, Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms, *Am. J. Bot.*, **87** (2000), 1712–1730.‡https://doi.org/10.2307/2656749

32. J. R. Bolton, I. Mayor-Smith, K. G. Linden, Rethinking the concepts of fluence (UV dose) and fluence rate: the importance of photon-based units–a systemic review, *Photochem. Photobiol.*, **91** (2015), 1252–1262. https://doi.org/10.1111/php.12512