International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 2471-2488

https://doi.org/10.70517/ijhsa464209

Analysis of the Application of Speech Recognition Technology in French Cross-Cultural Communication and Its Impact on Improving Students' Language Proficiency

Min Shang[1,2,*], Yifeng Wang[2] and Jin Chai[1]
[1] Xi'an International University, Xi'an, Shaanxi, 710077, China
[2] Xidian University, Xi'an, Shaanxi, 710126, China
Corresponding authors: (e-mail: isabelle9957@163.com).

Abstract This study focuses on the development of high-precision French speech recognition technology and its application in cross-cultural communication teaching. First, we propose an end-to-end French phoneme recognition method based on cross-modal knowledge distillation, using a CTC decoder to address phoneme alignment issues, and designing a frame-level distillation weight adaptation mechanism and sequence-level distillation. Additionally, we integrate speaker recognition technology based on i-vectors, using factor analysis to extract low-dimensional speaker features, thereby enhancing the system's adaptability to learners. We also propose a teaching strategy to enhance students' language proficiency by cultivating French thinking, creating authentic contexts, strengthening cross-cultural awareness, and establishing a layered interactive teaching model. Experiments based on French speech datasets show that the English pre-trained model performs optimally, with a CER of 8.87% and a SER of 10.46% between the Latin alphabet and the French alphabet set. The CTC decoder significantly outperforms the Transformer/Conformer, with a CER 9.42 percentage points lower than the Transformer encoder's 24.95%. After introducing i-vectors, the maximum error rate reduction reached 61.2%, and the syllable error rate SER on multilingual character sets decreased from 18.60% to 7.22%. Through stepwise multiple regression analysis of 476 student questionnaires, it was found that language attitude is the core predictor of conversational ability (β = 0.24, explaining 13.4% of the variance), self-efficacy dominates French proficiency improvement (β = 0.24, $\triangle R^2 = 0.065$), and learning resources contribute most to reading ability (β = 0.33, explaining 21.1% of the variance).

Index Terms French speech recognition, cross-modal knowledge distillation, CTC, i-vector, language ability

I. Introduction

As one of the most important international languages in the world, French has a significant global influence that cannot be overlooked. With the development of global economic trade and the advancement of globalization, cross-cultural communication between people of different linguistic and cultural backgrounds has become increasingly frequent. The demand for business cooperation and services conducted in French has been growing year by year [1], [2]. Therefore, the need to learn French has also been steadily increasing. In the new era, there is a severe shortage of composite talents who are proficient in foreign languages, understand their own culture, and can communicate smoothly with people from different cultural backgrounds despite cultural barriers [3]. Currently, traditional French language teaching processes exhibit issues such as outdated teaching methods and outdated teaching content, leading to some students having low motivation to learn, weak cross-cultural communication awareness, insufficient contextual adaptability, and inadequate cultural empathy, which severely hinder the improvement of students' language proficiency [4]-[6]. From the early stages of French learning, students encounter verb conjugation and gender distinctions in nouns, with many challenging concepts potentially deterring them and causing feelings of anxiety or frustration, which may lead to a loss of interest in French learning over time [7], [8]. Meanwhile, education is increasingly moving toward digitalization, and the French Ministry of Education has explicitly identified speech recognition technology as a key area for development in digital campuses.

Speech recognition technology is a major hot topic in the field of artificial intelligence in recent years. It refers to the technology of identifying the content of human speech from speech signals and converting it into text or commands that computers can understand. An increasing number of speech recognition technologies can achieve multilingual recognition, and the fields they cover are also expanding, with widespread application in areas such as speech translation, speech recognition navigation, smart home control, and intelligent voice assistants [9]-[11]. Speech recognition technology is playing an increasingly important role in cross-cultural communication. Through artificial intelligence technology, intelligent voice assistants can more accurately understand the content of human voice input and provide corresponding responses, assisting human translators in interpretation, eliminating subjective factors, and improving translation quality, thereby promoting cross-cultural communication [12], [13]. Literature [14] points out that speech recognition technology helps reduce cross-language communication barriers, primarily by improving communication effectiveness, reducing misunderstandings in cross-cultural communication, and quickly converting speech into text, thereby facilitating cooperation and communication in fields such as business and healthcare.

With the development of technologies such as machine learning and deep learning, the recognition rate and accuracy of speech recognition technology have been significantly improved. In particular, recent advancements in deep learning technology, through large amounts of training data and optimized algorithms, can achieve very high accuracy rates, even surpassing human recognition capabilities. Literature [15] utilizes deep neural network models combined with phonetic features to optimize the performance of speech recognition systems, enabling the recognition of phonetic features across different dimensions. Literature [16] designed a WAR strategy optimization algorithm to optimize spectral and statistical features, as well as machine learning model parameters, thereby obtaining a new classifier. This method achieves a speech recognition accuracy rate of up to 100% and a response speed as fast as 0.001 seconds.

Furthermore, with the improvement of computer computing power, speech recognition algorithms can achieve increasingly efficient operations, providing strong support for language learning [17]. Literature [18] analyzed the use of automatic speech recognition technology on two language learning websites, finding that speech recognition improved students' receptive vocabulary levels. Literature [19] showed that automatic speech recognition technology helped students standardize their spoken language and speak more fluently, although there was no significant improvement in spoken vocabulary and morphological grammar accuracy, which remained stable. Literature [20] demonstrates that in English and French instruction, speech recognition technology can create authentic spoken language practice scenarios and provide feedback, though it does not yet enable automatic correction and calibration of spoken language feedback. Literature [21] employs artificial intelligence speech recognition technology to design a French spoken language corpus, thereby exploring code-switching in teachers' discourse in French classrooms, offering insights for spoken language instruction. Literature [22] evaluated the effectiveness of Google Translate's text-to-speech and automatic speech recognition functions in pronunciation practice for French as a second language, with these functions enhancing language connection performance before and after pronunciation practice. Literature [23] used automatic speech recognition technology to understand French as a second language learners' self-awareness of clarity, and this technology can also clearly recognize the conditions required for interpersonal communication.

This study focuses on developing high-precision speech recognition technology suitable for French cross-cultural communication scenarios and exploring strategies for effectively integrating it into teaching to promote students' language development. First, we propose and discuss in detail a French automatic recognition method based on cross-modal knowledge distillation. The core of this method lies in using a trained complex teacher model to guide the training of a lightweight student model, thereby maintaining high precision while improving efficiency. A CTC decoder is employed to address the challenge of aligning speech frames with phoneme labels, enabling end-to-end prediction of French phoneme sequences. Two distillation methods—frame-level and sequence-level—are designed, with optimization achieved through loss functions such as KL divergence, cross-entropy (CE), and cosine similarity. To address the low efficiency of teacher knowledge transfer in traditional frame-level distillation, we innovatively propose a frame-level distillation weight adaptation method. This method automatically learns the correlation coefficients between tasks through a multi-task joint loss function, dynamically balancing the gradients of KL divergence and CE, effectively resolving the distillation failure issue caused by imbalanced loss gradients. Considering the characteristics of sequence models, we propose sequence-level distillation, using cosine similarity as the distillation loss to assist CTC loss, thereby avoiding the issue of KL divergence disrupting the CTC alignment mechanism, enabling the student model to learn sequence-level knowledge from the teacher model. Additionally, speaker recognition technology based on i-vectors is introduced. This technology, built on the GMM-UBM framework, uses factor analysis to map high-dimensional speaker Gaussian mean ultra-vectors into a low-dimensional global difference space, extracting compact i-vector features that encapsulate core speaker information. Finally, based on the precise feedback provided by the optimized speech recognition system, a systematic teaching strategy for improving students' language proficiency is proposed, including cultivating French thinking patterns, creating authentic language environments, fostering cross-cultural communication awareness, and transforming teachers' teaching concepts to establish a student-centered, tiered interactive teaching model.

II. Speech recognition and language ability training strategies based on knowledge distillation and i-vectors

II. A. French automatic recognition method based on cross-modal knowledge distillation

II. A. 1) Phoneme sequence decoding

To decode the target phoneme sequence from the audio semantic information, the model uses a CTC decoder. CTC trains the model end-to-end, establishing a many-to-one mapping between the learned continuous audio feature sequence and the sequence labels. In the field of speech recognition, it is extremely difficult to align labels with speech frames. However, CTC does not need to consider label alignment issues and is commonly used in network training in the field of speech recognition.

Let $x=\left(x^1,x^2,\ldots,x^T\right)$ be an audio time series, where each time slice is an audio feature vector $x^t=\left(x_1^t,x_2^t,\ldots,x_m^t\right)$. Let y be the corresponding label, whose length is less than or equal to T. The network's input is the power-normalized spectrum of the audio segment, and its output is a sequence of T length clue phonemes $y'=\left(y'^1,y'^2,\ldots,y'^T\right)$, where for each time step t, $y'^t=\left(y_1'^t,y_2'^t,\ldots,y_n'^t\right)$. At each time step t, the LSTM predicts phonemes, with the prediction probability being $p(y'^t\mid x)$. Here, $y'^t\in L$, where $L=\{blank,\Gamma\}$ where blank represents an empty audio segment at the current column, and Γ is the set of reference phonemes, which differ between French and British French. For each time step t, the length n of the prediction vector y'^t is equal to the length of the set L. The operation of removing duplicate characters and blank characters (e.g., '-') from the output sequence in the LSTM is defined as the B transformation. For example, when the input sequence length is $T=12$, the B transformation is:

$$B\left(\pi\right)=B\left(-sstta-t-e\right)=state \tag{1}$$

Among them, π represents an output sequence of the LSTM network. When optimizing LSTM, only the following probabilities need to be optimized:

$$p\left(y\mid x\right)=\sum_{B(\pi)=y}p\left(\pi\mid x\right) \tag{2}$$

That is, maximize the sum of the path probabilities that can obtain y. The CTC loss function is expressed as $L\left(x,y;\theta\right)$, where $\left(x,y\right)$ is a set of paired inputs and outputs, and θ is the current parameter of the network. The network parameters are updated by minimizing the negative of the loss function, i.e., minimizing the gradient of the loss function $\nabla_\theta L\left(x,y;\theta\right)$ of the loss function, i.e., by minimizing the negative value of equation (2).

II. A. 2) Knowledge distillation strategy

The knowledge distillation task consists of two sub-tasks: distillation from the teacher model to the student model and guidance from the true labels to the student model. For the distillation sub-task, the paper considers two distillation strategies: frame-level and sequence-level distillation. In frame-level distillation, the paper proposes a weight-adaptive loss function to balance the loss gradients between the two sub-tasks, thereby enhancing the effectiveness of distillation. In sequence-level distillation, the paper proposes using cosine similarity as the distillation loss to assist in joint training with CTC loss. Several basic loss functions are involved, including KL divergence, cross-entropy loss, cosine similarity, and CTC. CTC has been introduced in the previous section, and this section primarily introduces cross-entropy loss, KL divergence, and cosine similarity.

In information theory, probability p is a measure of certainty, while information is a measure of uncertainty. The amount of information is inversely proportional to the probability of an event occurring. Suppose X is a discrete random variable with a value set X and a probability distribution function $p(x)=P(X=x)$. $x\in\chi$, then the information content of the event $X=x_0$ is $I(x_0)=-\log(p(x_0))$. The relationship between entropy and information content is:

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log(p(x_i)) \tag{3}$$

For the 0-1 problem, i.e., the binary classification problem, it can be simplified to:

$$H(X)=-p(x)\log(p(x))-(1-p(x))\log(1-p(x)) \tag{4}$$

KL divergence, also known as relative entropy, is used to measure the difference between two probability distributions with the same variable x. It is a non-symmetric method for calculating the distance between two probability distributions $p(x)$ and $q(x)$, where p is often used to represent the true distribution. Specifically, the KL divergence between $q(x)$ and $p(x)$ is denoted as $D_{KL}(p(x),q(x))$, which measures the information loss when $q(x)$ approximates $p(x)$. Let $p(x)$ and $q(x)$ be two probability distributions of the discrete random variable x, i.e., $\sum_{x\in X} p(x)=1$, $\sum_{x\in X} q(x)=1$, and for every x, $p(x)>0$ and $q(x)>0$. The definition of $D_{KL}\left(p(x)\|q(x)\right)$ is shown in formula (5):

$$D_{KL}\left(p(x)\|q(x)\right)=\sum_{x\in X} p(x)\ln\frac{p(x)}{q(x)} \tag{5}$$

Among them, the smaller the value of D_{KL}, the closer the two distributions are.

Cross-entropy is closely related to KL divergence, and it transforms KL divergence. Its formula can be expressed as:

$$\begin{aligned} D_{KL}\left(p(x)\|q(x)\right) &= \sum_{i=1}^{n} p(x_i)\log(p(x_i))-\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \\ &= -H(p(x))+\left[-\sum_{i=1}^{n} p(x_i)\log(q(x_i))\right] \end{aligned} \tag{6}$$

$H(p(x))$ is the entropy of p, and the latter part is the cross-entropy:

$$H(p,q)=-\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{7}$$

For classification problems, p can be regarded as the true distribution of labels, so the entropy $-H(p(x))$ of p in the KL divergence is a fixed value, and the variable part is the cross-entropy. The purpose of neural network training is to make the predicted distribution q approximate the true distribution p, and cross-entropy can be used to replace KL divergence as the loss function. Cross-entropy and KL divergence are both widely used loss functions.

Cosine similarity calculates the similarity between vectors using the cosine value of the angle between them, primarily considering the consistency of the directions of the two vectors. When the angle is 0 degrees, meaning the two vectors are in the same direction, the cosine value is 1; when the angle is 180 degrees, meaning the two vectors are in opposite directions, the cosine value is -1; when the angle is 90 degrees, meaning the two vectors are perpendicular, the cosine value is 0. The larger the cosine similarity, the more consistent the directions of the two vectors.

II. A. 3) Frame-level distillation weight adaptive method

In frame-level distillation, traditional distillation methods enable student models to learn soft labels from teacher models. Soft labels are the outputs of a widely applied Softmax function, as shown in Formula (8), i.e.,

$$q_i=\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{8}$$

Among these, z_i is the network output before the Softmax function, q_i is the classification probability relative to each z_i, and T is the temperature coefficient. The larger T is, the smoother the distribution of soft labels becomes.

To improve the recognition accuracy of the student model, both the soft labels from the teacher model and the true labels are crucial. Therefore, we simultaneously consider the KL divergence between the probability distributions of the teacher model and the student model, and the cross-entropy loss between the probability distribution of the student model and the true labels, where the KL divergence serves as the distillation loss and the cross-entropy as the student loss. The joint loss function is shown in Formula (9):

$$L\left(x_t,x_s;W\right)=\alpha KL\left(P_{(T)},Q_{(T)}\right)T^2+\left(1-\alpha\right)CE\left(Y,Q_{(1)}\right) \tag{9}$$

Among these, x_t is the audio input of the cue text from the pre-trained teacher model, x_s is the fused feature vector, i.e., the input to the student model, W is the parameter of the student model, and α is a hyperparameter. $P_{(T)}$ and $Q_{(T)}$ are the Softmax outputs of the teacher model and student model, respectively, when the temperature parameter is T. Y is the true label, and $Q_{(1)}$ is the Softmax output of the student model when the temperature is 1.

For this traditional joint loss function, the values of T and α are manually adjusted to find a better student model. When T is set to 1 and α is set to 0.5, the loss curve of the joint loss of KL divergence and CE is shown in Figure 1. The gradient of the KL divergence is almost zero, indicating that the teacher model's knowledge has not been effectively distilled to the student model, and the gradients between the KL divergence and CE are unbalanced.

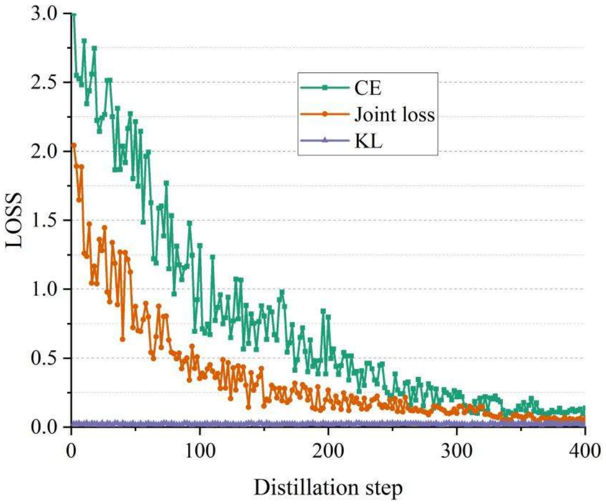

Figure 1: The loss curve of the combined loss of KL divergence and CE

Adjusting the values of T and α to obtain the optimal student model involves a multi-objective optimization problem, which is common in many deep learning problems. A simple method for combining multi-objective losses is to perform a weighted linear sum of each loss:

$$L_{total} = \sum_{i} w_i L_i \tag{10}$$

However, the model's performance is highly sensitive to the choice of weights, and adjusting weight hyperparameters is both time-consuming and labor-intensive, with no guarantee of obtaining the optimal model. To address the issue of unbalanced loss gradients between the two subtasks in knowledge distillation while avoiding the need for manual parameter tuning, the paper proposes a method using an adaptive weight loss function. This method first uses a multi-task joint loss function to automatically learn the correlation coefficients between tasks, and then uses these coefficients to calculate the balancing coefficients for balancing the loss gradients between the two sub-tasks. The derivation of the multi-task joint loss function is based on the uncertainty theory in multi-task learning, where uncertainty between tasks is inherent. Therefore, the knowledge distillation process can be viewed as two classification tasks, with the covariance uncertainty theory serving as the foundation for the weights between losses in knowledge distillation. The paper derives the joint loss function for the dual classification task by maximizing the Gaussian likelihood, with the derivation process as follows:

Let $f^W(x)$ be the output of the neural network, and W be the model parameters. The probability model for the classification task is defined as shown in formula (11):

$$p\left(y \mid f^W(x)\right) = Soft\max\left(f^W(x)\right) \tag{11}$$

The probability model for multiple tasks is shown in formula (12):

$$p\left(y_1,\ldots,y_K \mid f^W(x)\right) = p\left(y_1 \mid f^W(x)\right)\cdots p\left(y_K \mid f^W(x)\right) \tag{12}$$

Among them, $y_1,\ldots,y_K$ are model outputs.

Consider the classification problem as a re-scaling operation of the network output through Softmax:

$$p\left(y \mid f^{W}(x),\sigma\right)=Soft\max\left(\frac{1}{\sigma^{2}}f^{W}(x)\right) \tag{13}$$

σ is the noise parameter. σ^2 can be regarded as the Boltzmann constant kT and can be fixed or learned. According to the Boltzmann distribution, the maximum likelihood estimate for the classification problem is:

$$\log p(y=c \mid f^{W}(x),\sigma)=\frac{1}{\sigma^{2}}f_{c}^{W}(x)-\log\sum_{c'}\exp\left(\frac{1}{\sigma^{2}}f_{c'}^{W}(x)\right) \tag{14}$$

Among them, $f_c^W(x)$ is the c th element in the vector $f^W(x)$.

The joint loss function for the binary classification task is derived as follows:

$$\begin{aligned}L\left(W,\sigma_1,\sigma_2\right)&=-\log p\left(y_1,y_2=c \mid f^{W}(x)\right)\\&=Soft\max\left(y_1=c;f^{W}(x),\sigma_1\right)\cdot Soft\max\left(y_2=c;f^{W}(x),\sigma_2\right)\\&=-\log p\left(y_1=c \mid f^{W}(x),\sigma_1\right)-\log p\left(y_2=c \mid f^{W}(x),\sigma_2\right)\\&\approx\frac{1}{\sigma_1^{2}}L_1(W)+\frac{1}{\sigma_2^{2}}L_2(W)+\log\sigma_1+\log\sigma_2\end{aligned} \tag{15}$$

Among them, $L_1(W)$ is the KL divergence, and $L_2(W)$ is the cross-entropy loss.

Therefore, the multi-task joint loss function formula is:

$$L_{mtl}=\frac{1}{\sigma_1^{2}}KL\left(P,Q\right)+\frac{1}{\sigma_2^{2}}CE\left(Y,Q\right)+\log\sigma_1+\log\sigma_2 \tag{16}$$

This process does not require consideration of the two parameters T and α, but only needs to automatically learn the two observation noise parameters σ_1 and σ_2.

Furthermore, using the task-related coefficients σ_1 and σ_2 automatically learned from formula (16), the balance coefficient a shown in formula (17) can be obtained to solve the problem of gradient imbalance between the two subtasks mentioned above, while avoiding manual parameter tuning.

$$L\left(x_t,x_s;W\right)=\frac{1}{2}\left(a\times KL\left(P_{(1)},Q_{(1)}\right)+CE\left(Y,Q_{(1)}\right)\right) \tag{17}$$

Specifically, to obtain a, calculate $w_1=\frac{1}{\sigma_1^2}$ and $w_2=\frac{1}{\sigma_2^2}$ as prior knowledge, and then obtain the balance coefficient $a=w_1/w_2$.

II. A. 4) Sequential distillation

Frame-level knowledge distillation has limitations for sequence models because the sequence-level criteria used in training the teacher model do not align with the frame-level criteria used during distillation. Theoretically, sequence-level distillation could yield better performance for sequence models.

Similar to frame-level distillation, for sequence-level distillation, both the soft labels and true labels obtained from the teacher model play a crucial role in improving the student model's performance. Therefore, after determining that the student loss is CTC, it is necessary to explore a distillation loss that can assist CTC in enabling the student model to perform better. Since the KL divergence causes the student model to learn the soft labels of the teacher model frame-by-frame, which disrupts the alignment of CTC and reduces model accuracy, the paper uses cosine similarity instead of KL divergence as the distillation loss. The sequence-level distillation joint loss function is shown in Formula (18):

$$L\left(x_t,x_s;W\right)=\frac{1}{2}\left(1-\cos\left(S_S,S_t\right)+CTC\left(Y,\log Q\right)\right) \tag{18}$$

Among them, x_i, x_S, W, Q, and the corresponding parameters in formula (9) have the same meanings. S_S and S_t are the outputs of the student model and teacher model before the Softmax layer, respectively. Y is the phoneme transcription label corresponding to a sentence.

II. B.Speaker recognition based on i-vectors

The optimized speech recognition model can effectively decode French phonemes under ideal conditions. However, in real-world language learning scenarios, the system must contend with the varying pronunciation characteristics and speech quality of different learners (speakers). To enhance the system's robustness and adaptability to different speakers, speaker recognition technology based on i-vectors has been introduced.

The i-vector-based speaker recognition model is an improved version of the GMM-UBM model. However, unlike the GMM-UBM model, which trains a GMM model for each different speaker, the i-vector-based speaker recognition algorithm extracts i-vector features from each registered speech sample. These features contain all speaker information from the speech sample and can be used to characterize the target speaker.

II. B. 1) Speaker recognition algorithm based on GMM-UBM

Since the introduction of speaker recognition algorithms based on the GMM model, it has become the most widely used statistical model in text-independent speaker recognition. As a statistical model, the GMM model can be applied to speaker recognition based on the premise that each speaker's features follow a probability distribution that is identical in form but differs in parameters. GMM can fit any probability distribution, where each Gaussian probability function fits the feature distribution corresponding to different phonemes in the target speaker's speech signal. The mixture Gaussian model can be expressed as follows:

$$p(x \mid \lambda_s) = \sum_{i=1}^{M} \alpha_i^s b_i^s(x) \tag{19}$$

In the equation, x represents the D-dimensional feature vector; α_i^s denotes the mixture weights, satisfying $\sum_{i=1}^{M} \alpha_i^s = 1$, and $b_i^s(x)$ denotes the subdistribution. Each subdistribution is a D-dimensional joint Gaussian probability distribution, which can be expressed as:

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i^s\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_i^s\right)' \left(\Sigma_i^s\right)^{-1} \left(x-\mu_i^s\right)\right\} \tag{20}$$

In the equation, μ_i^s is the mean vector, and Σ_i^s is the covariance matrix. The complete mixture Gaussian model consists of the mean vector, covariance matrix, and mixture weights parameters. The mixture Gaussian model for speaker s is expressed as:

$$\lambda_s = \left\{\alpha_i^s, \mu_i^s, \Sigma_i^s\right\}, i = 1, \ldots, M \tag{21}$$

In practical applications, the amount of speech data available for training the target speaker is often very limited. When there is insufficient speech data, the Gaussian component count of the GMM model trained is very low, making it unable to fully capture the target speaker's characteristics. Therefore, Reynolds introduced UBM based on GMM. UBM is a GMM with a very large Gaussian component count, trained using a large amount of speech data unrelated to the target speaker. Since it is trained using a large amount of unrelated data, the UBM can be considered a background GMM that represents the distribution of all speech. All GMM models of registered target speakers are obtained from this background using an adaptive algorithm. The UBM model effectively solves the problem of incomplete models caused by insufficient registration data, thereby effectively improving recognition rates. At the same time, since the UBM model only needs to be trained once, and all registered speakers are obtained from this UBM through adaptation, it also effectively reduces the computational load.

The parameter estimation of GMM and UBM is based on the maximum likelihood criterion. Starting from a random or specific initial parameter λ, the EM algorithm is iterated so that the new model likelihood under each newly estimated parameter $\hat{\lambda}$ is not less than the previous value, i.e., $p(x \mid \hat{\lambda}) \geq p(x \mid \lambda)$. The EM iterative algorithm is primarily divided into two parts: the E step and the M step, with the formulas as follows:

(1) E step

According to Bayes' theorem, calculate the posterior probability of the latent variable a_j:

$$\hat{\gamma}_{ij} = \frac{\alpha_i \phi\left(x_j \mid \lambda_i\right)}{\sum_{i=1}^{M} \alpha_i \phi\left(x_j \mid \lambda_i\right)}, i = 1,2,\ldots,M; j = 1,2,\ldots,N \tag{22}$$

(2) Step M
Update the parameters in the model.

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{N} \hat{\gamma}_{ij}}{N}, i = 1,2,\ldots,M \tag{23}$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{N} \hat{\gamma}_{ij} x_j}{\sum_{j=1}^{N} \hat{\gamma}_{ij}}, i = 1,2,\ldots,M \tag{24}$$

$$\hat{\Sigma}_i^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{ij} (x_j - \mu_i)^2}{\sum_{j=1}^{N} \hat{\gamma}_{ij}}, i = 1,2,\ldots,M \tag{25}$$

Repeat the above calculations until the log-likelihood function no longer shows significant changes.

After training the UBM using a large amount of irrelevant data, it is necessary to use the registered data of the target speaker to obtain the model of each speaker on the UBM through an adaptive algorithm. The adaptive algorithm adopts the maximum a posteriori (MAP) criterion, whose basic idea is to update the parameters of the Gaussian component in the UBM so that the parts of the parameters related to the target speaker are fitted in its direction, while the remaining parts remain unchanged. The MAP speaker adaptation algorithm is shown in Figure 2:

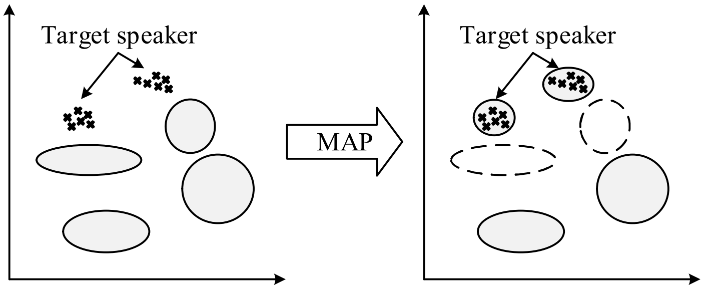

Figure 2: MAP Adaptive Algorithm

In the target speaker model obtained using registered speech and the MAP adaptive algorithm, due to the sparsity of the registered speech, only some of the parameters of the Gaussian components in the UBM are adjusted, while the noise information of the test speech is the same in both the speaker model and the UBM. The test speech features calculate the probability of each speaker model appearing using the log-likelihood ratio, subtract the corresponding probability in the UBM, and obtain the final score for each speaker, as shown in the following formula:

$$S(O) = \log P(O \mid \lambda_{GMM}) - \log P(O \mid \lambda_{UBM}) \tag{26}$$

II. B. 2) Speaker recognition algorithm based on i-vectors

The MAP-adaptive GMM-UBM algorithm effectively addresses the issue of sparse target speaker registration in speaker recognition. Experiments show that the mean values of the Gaussian mixtures in the GMM-UBM model contain nearly all speaker information, so extracting the mean values of all Gaussian components to form a mean supervector can effectively classify speakers. However, the mean supervector has a very high dimension. Therefore, N. Dehak proposed an i-vector-based model for speaker modeling based on the global difference space. The i-vector-based speaker recognition algorithm has become the most widely used model today.

The i-vector-based speaker recognition algorithm builds upon the GMM-UBM mean supervector by employing factor analysis to propose a global difference space, which incorporates both channel effects and speaker information from the speech. The mean supervector is projected onto a low-rank global difference space, and through this dimensionality reduction operation, a low-dimensional vector containing only speaker information is obtained, which becomes the i-vector. This is expressed mathematically as follows:

$$m_s = m_0 + T\omega \tag{27}$$

In the equation, m_s denotes the mean supervector in the speaker's GMM-UBM model; m_0 denotes the mean supervector of the UBM, which is independent of both the speaker and the channel; T denotes the global difference space; and ω is the global variation factor, i.e., the low-dimensional vector i-vector. Before calculating the i-vector features, the global difference space matrix T must be estimated. Similar to the GMM and UBM methods described earlier, the EM algorithm is used to estimate T. After obtaining the global difference space matrix, the i-vector features of a segment of speech are extracted using the following formula:

(1) Calculate the Baum-Welch statistic:

$$N_c = \sum_{t=1}^{L} P(c \mid x_t) \tag{28}$$

$$F_c = \sum_{t=1}^{L} P(c \mid x_t) x_t \tag{29}$$

$$\tilde{F}_c = \sum_{t=1}^{L} P(c \mid x_t)(x_t - m_c) \tag{30}$$

In the equation, x_t denotes the D-dimensional MFCC feature vector of the tth frame of speech, L is the length of the speech, c denotes the Gaussian component, $P(c \mid x_t)$ denotes the posterior probability that x_t belongs to the Gaussian component c, and m_c is the mean vector of the Gaussian component c. N_c is the zero-order statistic, F_c is the first-order statistic of dimension $D \times 1$, and $\tilde{F}_c$ is the first-order central statistic of dimension $D \times 1$.

(2) Extracting i-vector features

$$\omega = \left(I + T^t \Sigma^{-1} N T\right)^{-1} T^t \Sigma^{-1} \tilde{F} \tag{31}$$

where I is the identity matrix, T is the global difference space matrix estimated from a large amount of data, Σ is the covariance matrix of UBM, N is the square matrix of the CD×CD dimension composed of N_c living diagonal elements, and $\tilde{F}$ is the vector of the $CD \times 1$ dimension composed of $\left(\tilde{F}_c\right)$.

After extracting i-vector features for all registered speech samples, a model can be constructed for the registered speakers. Since the i-vector features already contain all speaker features, this paper calculates the average of all i-vector features for the same speaker to obtain a mean i-vector vector representing that speaker. During the testing phase, i-vector features are extracted from the test speech using the same steps, and scores are calculated by matching them with the mean i-vector vectors of all registered speakers. This paper uses cosine distance scoring, with the highest score indicating that the test speech corresponds to that speaker.

II. C.Teaching Strategies for Improving Students' Language Proficiency

By optimizing the core recognition engine through cross-modal knowledge distillation and combining it with i-vector technology to enhance adaptability to individual learners, we have built a high-precision, robust speech recognition system for French language teaching. This chapter systematically elaborates on targeted teaching strategies for improving students' language proficiency from four aspects.

II. C. 1) Cultivating students' French way of thinking

The cultivation of students' French thinking patterns involves encouraging them to think in French and minimizing the negative transfer from their native language. In the classroom, when students express themselves in French, teachers should consciously cultivate and guide them to think directly in French, minimizing the influence of their native language. Language is a tool, and mastering it is a process of practice and repetition. With persistence and extensive practice, students will inevitably become proficient and able to use it flexibly. Outside of class, teachers require students to read foreign newspapers, magazines, and books, as well as analyze foreign films and TV shows. They should learn to accumulate authentic sentences, dialogues, and colloquial expressions from these sources.

In this way, students' French thinking patterns are cultivated and enhanced without them even realizing it.

II. C. 2) Creating a French language learning environment

The environment has a significant impact on language learning. In the classroom, teachers should make every effort to create an authentic French-language learning environment and teach in French to develop and enhance students' language skills. Outside of class, schools should organize more French-language activities such as “French Corners” and “French Culture Festivals,” allowing students to immerse themselves in a foreign language and cultural environment, thereby understanding and mastering French in a subtle and gradual process. At the same time, schools can also offer more interest-based classes and hire foreign teachers to enable students to interact with them, thereby applying French in a real-life language environment.

II. C. 3) Cultivating cross-cultural communication awareness

The study of language knowledge is inseparable from cultural learning, as culture and language complement each other. In teaching, educators should consciously integrate knowledge of British and American culture into their lessons. By comparing Chinese culture with foreign cultures, particularly British and American cultures, students can gain a deeper understanding. Additionally, educators should create authentic real-life scenarios and encourage group collaboration, using French for communication, to engage in dynamic classroom activities. Educators should provide appropriate guidance, while students, immersed in this experiential learning environment, can develop a deeper reflection on the distant French culture. Outside of class, teachers can recommend classic and appropriate foreign films for students to analyze and require students to listen to French classic songs regularly, consciously cultivating their cross-cultural awareness to enhance their French language proficiency.

II. C. 4) Changing Teaching Concepts

University French teachers should promptly adjust their roles, shift their teaching philosophies, and establish a new type of equal teacher-student relationship that centers on the student. Teachers should transition from being knowledge transmitters to facilitators of student learning, emphasizing the student's central role and promoting their comprehensive development. Language acquisition cannot be achieved overnight but requires sustained, daily practice over an extended period. In this process, teachers must effectively assume the role of guides, tailoring instruction to individual student differences in French proficiency, learning habits, and other factors. They should adopt a differentiated, tiered teaching approach to replace the previous one-size-fits-all method, uncover students' potential, meticulously design instructional activities, enhance teacher-student interaction, encourage active student participation in class, and guide them in actively constructing French knowledge, thereby enhancing their language proficiency.

III. Research on speech recognition based on knowledge distillation and i-vectors

After proposing a speech recognition technology framework and teaching strategy based on cross-modal knowledge distillation and i-vectors, this study designed three comparative experiments to verify its actual effectiveness. First, it explored the impact of different source language pre-training on French phoneme recognition. Second, it compared the performance differences between CTC, Transformer, and Conformer encoders. Finally, it verified the improvement in system robustness after introducing i-vector technology.

III. A. Experimental setup

III. A. 1) Pre-training settings

The pre-training fine-tuning method primarily assists French training by inheriting model parameters. First, French text and source language text are preprocessed according to the corresponding modeling primitives and source language requirements. Next, the source language data is used as the training dataset to fully train the model, and the model parameters are saved. Finally, the saved model parameters are substituted for the randomly initialized model parameters in the French training model, and the French data is trained on this basis to fine-tune the parameters, resulting in the final French speech recognition model.

Using CTC as the end-to-end model, the number of hidden nodes in the gated layer is 256, the number of hidden units in the residual connections is also 256, the learning rate is 1e-4, and the optimizer is Adam. All models use batch training, trained for 500 epochs on a Linux system with two Nvidia RTX 2070 Super CPUs, with a batch size of 100.

III. A. 2) Evaluation Indicators

The evaluation metrics of a speech recognition system can reflect the current performance of the system. Depending on the modeling unit and application scenario, the main metrics include character error rate (CER), word error rate (WER), and sentence error rate (SenER). This paper focuses on French. In order to more accurately reflect the results of speech recognition, this paper uses syllable error rate (SER) as the basis for judging the performance of the system. The formula for calculating the syllable error rate is as follows:

$$SER = \frac{S+D+I}{L} \times 100\% \tag{32}$$

In this equation, the denominator L represents the number of syllables in the test sequence, while the numerator denotes the edit distance between the model's recognition result sequence and the test sequence. Edit distance is a method used to measure the degree of difference between two strings, primarily involving the calculation of the minimum number of operations required to transform one string into another. These operations include substitution, deletion, and insertion, represented by S, D, and I, respectively, in Equation (32).

III. A. 3) Dataset

The study used the open-source Common Voice French subset as the main data set, supplemented by the self-built SPROUTS French dialogue corpus. The total duration is 65.2 hours, containing 35,848 audio samples, covering read texts (news, literary works) and real dialogue scenarios (shopping, academic discussions, social interactions).

The English data comes from the LibriSpeech ASR corpus. The text is sourced from public domain audiobooks from the LibriVox project, covering multiple fields such as literature and technology. The Chinese dataset is the open-source TIIGIS-30 dataset released by the Intelligent Speech Laboratory at Tsinghua University, containing 31.8 hours of French speech. The text is selected from news articles, with 98% of the speakers being university students fluent in French, ensuring standardized pronunciation.

III. B. Speech Recognition Comparison Results

III. B. 1) Comparative Study of Speech Recognition Using Different Training Source Language Models

The experiment compared four modeling primitives: a longitudinal Latin alphabet set, a multilingual character set, Chinese syllables, and French alphabet sets, as well as Latin and French alphabet sets. Horizontally, it compared different training data sets formed by selecting different source languages (Chinese pre-training, English pre-training, and Chinese-English pre-training). The speech recognition error rates of the CTC model based on the knowledge distillation method are shown in Table 1, and bar charts of speech recognition error rates under different source language pre-training models are shown in Figure 3.

Table 1: Error rate of speech recognition based on CTC model

Dataset	Type	Chinese pre-training	English pre-training	Chi-Eng pre-training
Set of Latin letters	CER	21.69	15.53	19.02
	SER	20.23	14.81	17.93
Set of Multilingual character	CER	26.02	20.42	22.43
	SER	18.60	11.66	15.64
Set of Chinese syllables and French letters	CER	24.01	12.72	18.14
	SER	25.91	10.44	14.09
Set of Latin letters and French letters	CER	14.28	8.87	11.39
	SER	16.14	10.46	14.74

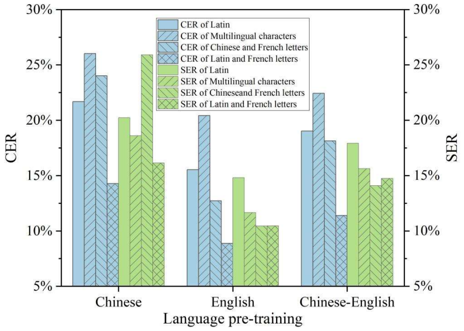

Figure 3: Error rates of recognition under different language pre-trained models

It can be seen that, among all modeling primitives, the character error rate (CER) and syllable error rate (SER) of English pre-training are both lower than those of Chinese pre-training. For example, in the Latin alphabet set, the CER of English pre-training is 15.53%, which is 6.16 percentage points lower than the 21.69% of Chinese pre-training. In the Latin alphabet and French alphabet datasets, the SER for English pre-training is 10.46%, which is 5.68 percentage points lower than the 16.14% for Chinese. The experimental results indicate that English is more suitable than Chinese as the source language for French as the target language. This is because there are more similarities in pronunciation between English and French. English and French belong to the Indo-European language family, while Chinese belongs to the Sino-Tibetan language family. Languages within the same language family can learn more similar knowledge in transfer learning.

III. B. 2) Comparative study of speech recognition based on different encoder models

To further validate the role of the CTC phoneme sequence decoder in French speech recognition, comparative experiments were conducted using Transformer encoders and Conformer encoders as baseline models. On the encoder side, the number of self-attention heads is 5, with a total of 10 self-attention modules. The feature dimension for each time step in the output is 286, the number of hidden nodes in the neural network within the attention layer is 691, and the convolution kernel size is 5×5. On the decoder side, there are no convolutional kernels, the total number of self-attention modules is 8, and the rest of the configuration is the same as on the encoder side. The models were pre-trained using English as the source language, and the speech recognition error rates obtained by the three models on different datasets are shown in Table 2 and Figure 4.

Table 2: The speech recognition error rates obtained on different datasets

Dataset	Type	CTC	Transformer	Conformer
Set of Latin letters	CER	15.53	24.95	20.87
	SER	14.81	21.11	17.86
Set of Multilingual character	CER	20.42	24.89	21.96
	SER	11.66	18.39	14.93
Set of Chinese syllables and French letters	CER	12.72	20.22	16.91
	SER	10.44	19.40	13.22
Set of Latin letters and French letters	CER	8.87	24.34	20.86
	SER	10.46	19.05	15.15

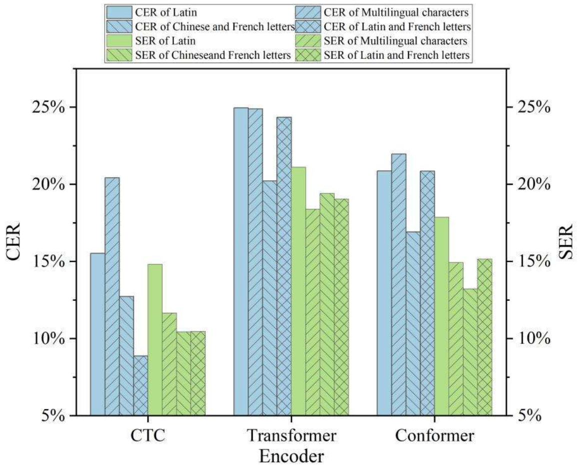

Figure 4: The speech recognition error rates obtained on different datasets

It can be seen that the CTC model demonstrates comprehensive advantages. Among the four modeling primitives, the CTC's CER and SER are both lower than those of the Transformer and Conformer. In the Latin alphabet set, the CTC's CER is 15.53%, which is 9.42 percentage points lower than the Transformer's 24.95%. In the Chinese syllable and French alphabet datasets, CTC's SER is 2.78 percentage points lower than Conformer's 13.22%. Additionally, CTC achieves the lowest SER across multilingual character sets at 11.66%, while Transformer and Conformer's SERs are as high as 18.39% and 14.93%, respectively, indicating that CTC is more suitable for handling multilingual mixed scenarios.

III. B. 3) Research on speech recognition based on i-vector-CTC

After verifying the advantages of the CTC decoder, we introduced speaker recognition technology based on i-vectors. The MAP-adaptive GMM-UBM algorithm effectively solved the problem of sparse target speaker speech in speech recognition. The speech recognition error rate of the CTC model after introducing i-vectors is shown in Table 3.

Table 3: The error rate of speech recognition of i-vector-CTC

		Chinese pre-training		English pre-training		Chi-Eng pre-training	
Dataset	Type	CTC	i-vector-CTC	CTC	i-vector-CTC	CTC	i-vector-CTC
Set of Latin letters	CER	21.69	14.32	15.53	12.44	19.02	13.42
	SER	20.23	13.24	14.81	11.08	17.93	12.22
Set of Multilingual character	CER	26.02	13.93	20.42	10.36	22.43	12.64
	SER	18.60	7.22	11.66	5.58	15.64	6.12
Set of Chinese syllables and French letters	CER	24.01	12.63	12.72	10.62	18.14	11.49
	SER	25.91	13.48	10.44	10.99	14.09	12.12
Set of Latin letters and French letters	CER	14.28	6.14	8.87	4.81	11.39	5.77
	SER	16.14	7.65	10.46	6.73	14.74	7.02

After introducing i-vector speaker recognition technology, the CTC model significantly reduced the error rate. Under English pre-training, the CER of the Latin alphabet and French alphabet sets decreased from 8.87% to 4.81%, a decrease of 45.8%, and the SER decreased from 10.46% to 6.73%, a decrease of 35.7%. Under Chinese pre-training, the SER for the multilingual character set decreased from 18.60% to 7.22%, a reduction of 61.2%. i-vector demonstrates particularly significant effectiveness in sparse speech scenarios (such as Chinese pre-training), with the CER for the Latin alphabet set decreasing from 21.69% to 14.32%. After combining Chinese-English pre-training with i-vector, the CER for the Latin alphabet set decreased to 13.42%, indicating that i-vector can effectively mitigate multilingual interference issues.

IV. Research on factors influencing students' language proficiency based on stepwise multiple regression

Experiments have demonstrated that the optimized speech recognition system significantly reduces syllable error rates; however, the ultimate value of the technology's effectiveness must be validated through educational practice. Therefore, this study further collected student questionnaire data and employed stepwise multiple regression analysis to explore the quantitative relationship between technology-supported teaching strategies and improvements in students' language proficiency.

This study selected undergraduate students majoring in French from the first to fourth years at a certain university as the research subjects. A questionnaire was distributed to students of this major via Questionnaire Star, with 500 questionnaires distributed and 476 valid questionnaires returned. Among these, 144 were from first-year students, 126 from second-year students, 114 from third-year students, and 92 from fourth-year students.

The variables “conversational language,” “French proficiency,” and “reading ability” were selected to reflect the French language proficiency of university students. Language behavior is often influenced by linguistic or non-linguistic factors. Therefore, the second part of the questionnaire investigated the factors influencing the French language proficiency of the study participants. Six variables were selected as explanatory variables: “parental education level,” “learning resources,” “self-efficacy,” “learning anxiety,” “health status,” and “language attitude.” Each dimension of the questionnaire was scored on a 5-point scale (1-very poor, 2-poor, 3-average, 4-good, 5-excellent).

This chapter will sequentially explain the results of descriptive statistical analysis, Pearson correlation analysis, and stepwise multiple regression analysis. Based on statistical data, it will analyze the current state of college students' French language proficiency, verify whether the factors involved in this study are correlated with college students' language proficiency, and examine the extent to which these factors influence college students' French language proficiency.

IV. A. Analysis of the Current State of French Language Proficiency Among College Students

Descriptive statistical analysis primarily involves statistical description and comparison of the distribution of various variables. The preliminary statistical results of the variables involved in this study are shown in Table 4. The specific scores of the student survey questionnaire for each dimension are plotted in a bar chart, as shown in Figure 5.

Table 4: Descriptive statistical analysis of each variable

Variable	M	SD	Max	Min
Conversation language	4.49	0.91	5	1
French proficiency level	4.15	1.21	5	1
Reading ability	3.76	1.23	5	1
Self-efficacy	3.92	1.48	5	1
Learning anxiety	3.45	1.44	5	1
Health condition	4.05	1.39	5	1
Language attitude	4.1	1.29	5	1
Parents' educational level	4.23	1.29	5	1
Learning resources	3.93	1.29	5	1

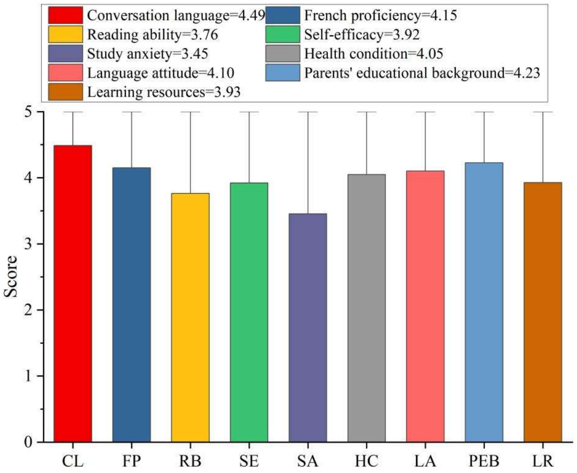

Figure 5: Specific scores of the student survey questionnaire in each dimension

The survey results indicate that the language proficiency of French language majors and related influencing factors are generally at an above-average level, with an overall mean range of 3.45–4.49. Among these, conversational language performance was the strongest, with a mean of 4.49, followed by French proficiency (M=4.15, SD=1.21), while reading ability was relatively weaker (M=3.76, SD=1.23). Parents' educational attainment (M=4.23, SD=1.29) and language attitude (M=4.10, SD=1.29) scored relatively high, while learning anxiety had the lowest mean (M=3.45, SD=1.44), indicating that students generally experience moderate anxiety.

IV. B. Correlation test

Before conducting stepwise multiple regression analysis, it is necessary to first test the correlation between the dependent variables of language ability (conversational language, French proficiency, reading ability) and the explanatory variables of influencing factors (parental education level, learning resources, self-efficacy, learning anxiety, health status, language attitude) in this study. Only when there is a significant correlation between the variables does it make sense to conduct regression analysis to find their specific form of correlation. The results of the Pearson correlation analysis conducted in this study are shown in Figure 6 below.

There is a significant correlation between language proficiency and influencing factors. Conversational language is strongly positively correlated with French proficiency (r = 0.578), language attitude (r = 0.474), and health status (r = 0.321); it is negatively correlated with learning anxiety (r = -0.113). French proficiency is highly correlated with self-efficacy (r=0.448) and language attitude (r=0.399). Reading ability is primarily positively correlated with learning resources (r=0.368) and self-efficacy (r=0.478), and negatively correlated with learning anxiety (r=-0.243).

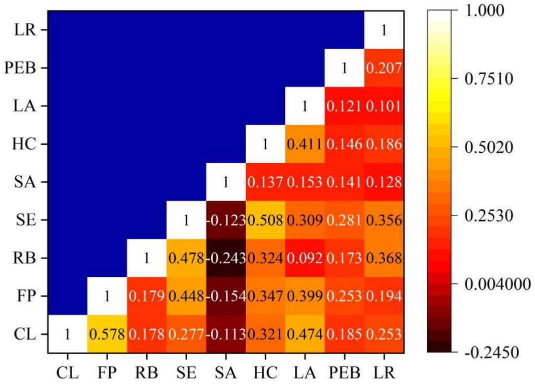

Figure 6: The results of Pearson correlation analysis

IV. C. *Stepwise multiple regression analysis of factors influencing conversational language*

To further explore the influence of parents' educational attainment, learning resources, self-efficacy, learning anxiety, health status, and language attitudes on college students' language proficiency, and to establish an “optimal” multiple linear regression model, this study conducted a stepwise multiple regression analysis. The influencing factors (parents' educational attainment, learning resources, self-efficacy, learning anxiety, health status, and language attitudes) were used as explanatory variables, and language proficiency (conversational language, Mandarin proficiency, and reading ability) as the dependent variables. A total of three regression models were established. To address the issue of high correlation among variables, this study conducted multicollinearity tests on the explanatory variables. The data showed that the variance inflation factor (VIF) values for all explanatory variables were less than 10, indicating no multicollinearity among the explanatory variables. This ensured the accuracy and reliability of the estimated correlation coefficients in the model and ruled out the possibility of spurious regression. The specific results are as follows.

IV. C. 1) Analysis of Factors Affecting College Students' Conversational Language

A stepwise multiple regression analysis was used to examine the predictive validity of parents' educational attainment, learning resources, self-efficacy, learning anxiety, health status, and language attitudes on college students' conversational language. The results are shown in Table 5.

Table 5: Multivariate Regression Analysis of Conversation Languages

Variable	B	SD	Beta	t	F	R^2	$\triangle R^2$
Constant	2.51	0.09	-	16.64***	146.29***	0.134	0.131
Language Attitude	0.66	0.14	0.39	12.16***			
Constant	2.08	0.12	-	11.81***	90.07***	0.142	0.141
Language Attitude	0.45	0.09	0.31	9.16***			
Learning Resources	0.31	0.05	0.25	7.94***			
Constant	1.67	0.16	-	9.46***	62.93	0.164	0.164
Language Attitude	0.38	0.13	0.27	8.54***			
Learning Resources	0.20	0.03	0.19	6.86***			
Self-Efficacy	0.16	0.14	0.14	5.55***			
Constant	1.35	0.04	-	8.70***	49.30***	0.191	0.191
Language Attitude	0.31	0.1	0.24	7.22***			
Learning Resources	0.17	0.15	0.15	6.54***			
Self-Efficacy	0.12	0.07	0.09	5.96***			
Parental Education Level	0.09	0.03	0.06	4.61***			

The stepwise regression model ($R^2 = 0.191$) indicates that language attitude is the core predictor of conversational language ($\beta = 0.24$), with significant predictor variables: language attitude ($\beta = 0.24$, $t = 7.22$), learning resources ($\beta = 0.15$, $t = 6.54$), self-efficacy ($\beta = 0.09$, $t = 5.96$), and parental education level ($\beta = 0.06$, $t = 4.61$). The four variables together explain 19.1% of the variance in conversational language, with language attitude alone contributing 13.4%.

IV. C. 2) Analysis of Factors Affecting College Students' French Proficiency Levels

A stepwise multiple regression analysis was conducted to examine the predictive validity of parents' educational attainment, learning resources, self-efficacy, learning anxiety, health status, and language attitudes on college students' French proficiency. The results are shown in Table 6.

Table 6: Multivariate Regression Analysis of French proficiency level

Variable	B	SD	Beta	t	F	R^2	△R^2
Constant	2.71	0.15	-	24.14***	172.35	0.166	0.165
Self-efficacy	0.71	0.06	0.43	12.31***			
Constant	2.15	0.13	-	15.08***	141.09	0.215	0.213
Self-efficacy	0.60	0.03	0.38	11.55***			
Language attitude	0.41	0.13	0.27	8.65***			
Constant	1.78	0.09	-	10.7***	86.63	0.238	0.234
Self-efficacy	0.45	0.04	0.26	8.76***			
Language attitude	0.31	0.12	0.20	7.14***			
Health condition	0.20	0.16	0.18	5.86***			
Constant	1.58	0.14	-	8.74***	69.04	0.256	0.255
Self-efficacy	0.40	0.15	0.24	7.33***			
Language attitude	0.30	0.13	0.17	7.06***			
Health condition	0.18	0.05	0.14	5.69***			
Parent's educational level	0.14	0.09	0.08	4.04***			
Constant	1.69	0.04	-	8.22***	53.82	0.269	0.267
Self-efficacy	0.39	0.08	0.24	7.04***			
Language attitude	0.28	0.08	0.17	6.85***			
Health condition	0.16	0.17	0.12	4.66***			
Parent's educational level	0.14	0.02	0.06	3.01***			
Learning anxiety	-0.06	0.12	-0.05	2.41**			

The five-variable regression model ($R^2 = 0.269$) indicates that the core driving factors are self-efficacy ($\beta = 0.24$, $t = 7.04$), language attitude ($\beta = 0.17$, $t = 6.85$), and health status ($\beta = 0.12$, $t = 4.66$). Learning anxiety has a weak negative impact ($\beta = -0.05$, $t = 2.41$). The introduction of self-efficacy increases the explanatory power by 6.4% (△R^2 = 0.065), highlighting its critical role.

IV. C. 3) Analysis of Factors Affecting College Students' Reading Ability

A stepwise multiple regression analysis was used to examine the predictive power of parents' educational attainment, learning resources, self-efficacy, learning anxiety, health status, and language attitudes on college students' reading ability. The results are shown in Table 7 below.

Table 7: Multivariate regression analysis of reading ability

Variable	B	SD	Beta	t	F	R^2	△R^2
Constant	1.01	0.12	-	6.29***	241.28***	0.211	0.209
Learning resources	0.69	0.05	0.50	16.39***			
Constant	0.78	0.12	-	4.11***	160.27***	0.278	20.277
Learning resources	0.49	0.02	0.39	13.04***			
Parent's educational attainment	0.37	0.03	0.27	6.21***			
Constant	1.36	0.09	-	6.67***	109.56***	0.288	0.285
Learning resources	0.45	0.15	0.33	12.74***			
Parent's educational attainment	0.31	0.04	0.20	6.02***			
Self-efficacy	0.22	0.14	0.12	9.59***			

The three-variable model ($R^2 = 0.288$) revealed that learning resources contributed the most to the dominant factors ($\beta = 0.33$, $t = 12.74$), followed by parental education ($\beta = 0.20$, $t = 6.02$) and self-efficacy ($\beta = 0.12$, $t = 9.59$).

Learning resources alone explain 21.1% of the variance in reading ability (first stage △R^2 = 0.209).

IV. D. Residual Normal Distribution and Homogeneity of Variance Test

In regression analysis, this study used regression standardized residual histograms and regression standardized residual scatter plots to determine whether the residuals were normally distributed and whether there was heteroscedasticity. The regression standardized residual histograms and regression standardized residual scatter plots are shown in Figures 7 and 8, respectively.

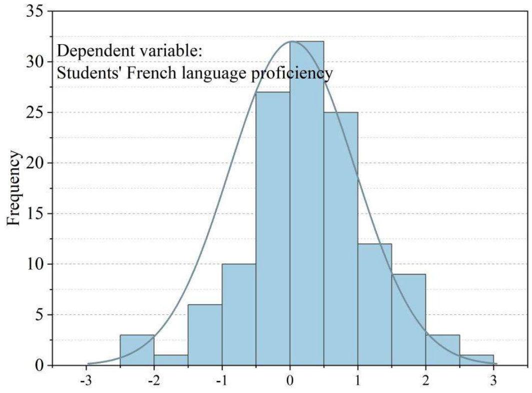

Figure 7: Regression standardized residual histogram

Figure 7 shows that the residuals in this study conform to a normal distribution, indicating good fitting results.

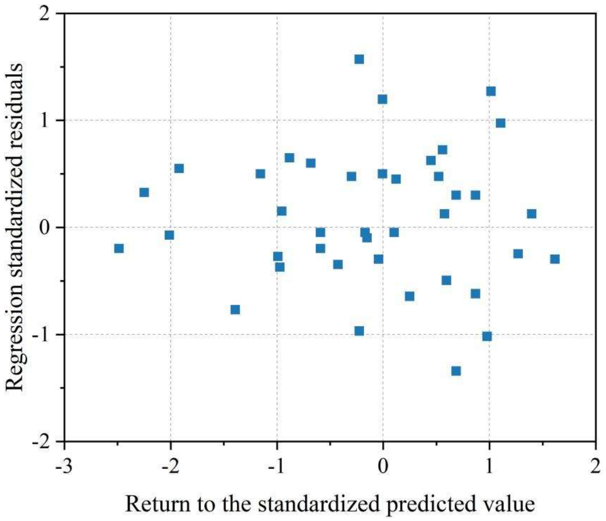

Figure 8: Regression standardized residual scatter plot

As shown in Figure 8, regardless of how the predicted values change within a specific range, the corresponding residuals always remain near the zero line, with the amplitude remaining basically stable and no obvious signs of heteroscedasticity. Therefore, it can be determined that there is no significant difference between the standardized residuals and the standardized normal distribution, which indicates that the conditions are met and the regression model is appropriate and feasible.

V. Conclusion

This study systematically verified the promotional effect of optimized speech recognition technology on improving French cross-cultural communication skills through a combination of technological innovation and teaching practice.

Experiments based on French speech data from multiple scenarios showed that the CTC model based on sequence-level distillation reduced the CER by 6.66 percentage points (from 15.53% to 8.87%) on the Latin alphabet set compared to traditional distillation. i-vector speaker recognition technology demonstrated outstanding optimization effects in multilingual mixed scenarios, with SER for multilingual character sets decreasing from 18.60%

to 7.22%, a reduction of 61.2%.

Stepwise multiple regression analysis based on 476 valid questionnaires confirmed that conversational ability is primarily driven by language attitude (β=0.24), with the four variables collectively explaining 19.1% of the variance; Improvements in French proficiency primarily depend on self-efficacy (β = 0.24), with an increase in explanatory power of 6.4% ($\triangle R^2 = 0.065$) after its inclusion; reading ability is mainly influenced by learning resources (β = 0.33), contributing 21.1% of the variance on its own.

References

[1] Oriakpono, M. (2024). The Role of the French Language in facilitating Trade: A Case Study of Maje Border Town Markets. NIU Journal of Humanities, 9(1), 213-223.

[2] Tijani, M. A. (2018). Globalization and the imperatives of French in Nigeria. Journal of Modern European Languages And Literatures, 1-8.

[3] Lifintsev, D., Zelihic, M., Grebliauskiene, B., Wellbrock, W., Patel, S. V., & Sharma, R. K. (2025). Young professionals' perspectives on cross-cultural communication: Assessing competence and employer support across regions. International Journal of Cross Cultural Management, 14705958251319695.

[4] Ogunbiyi, O., Makinde, S. O., & Ogunleye, O. R. (2024). French language teaching in the 21st Century; Challenges & Prospects. Fuoye Journal of Educational Management, 1(2).

[5] Lin, W. (2024). Research on the Current Situation of French Language Teaching in Chinese University Education. Open Access Library Journal, 11(12), 1-7.

[6] Agbor, C. A., & Ashabua, D. A. (2018). Basic education curriculum in South-South Nigeria: Challenges and opportunities of quality contents in French language learning. Global Journal of Educational Research, 17(2), 97-101.

[7] Myslihaka, L. (2017). Communication as the center of teaching/learning process of foreign languages (the case of French language). European Journal of Social Science Education and Research, 4(1), 1-11.

[8] Ismail, A. A., & Mohtar, W. I. W. (2024). Analysing Communicative Anxiety in French among French Language Students at the Faculty of Modern Languages and Communication, Universiti Putra Malaysia. Gading Journal for Social Sciences (e-ISSN 2600-7568), 27, 21-33.

[9] Baladari, V. (2023). Building an Intelligent Voice Assistant Using Open-Source Speech Recognition Systems. Journal of Scientific and Engineering Research.

[10] Ciobanu, D., & Secară, A. (2019). Speech recognition and synthesis technologies in the translation workflow. In The Routledge Handbook of Translation and Technology (pp. 91-106). Routledge.

[11] Ali, A. T., Eltayeb, E. B., & Abusail, E. A. A. (2017). Voice recognition based smart home control system. International Journal of Engineering Inventions, 6(4), 01-05.

[12] Kamegne, Y., Owusu, E., & Oladapo, H. (2025). Alexa, Why Can’t You Hear Our Accents: Cross Cultural Studies on the Inclusivity of Voice Recognition Systems. In International Conference on Human-Computer Interaction (pp. 62-71). Springer, Cham.

[13] Jing, C., & Liu, G. (2022). Acquisition of english corpus machine translation based on speech recognition technology. Scientific Programming, 2022(1), 5617400.

[14] Dai, R. Q. (2024). Overcoming cross-language communication barriers with speech recognition technology. Applied and Computational Engineering, 74, 53-58.

[15] Guan, W. (2018). Performance optimization of speech recognition system with deep neural network model. Optical Memory and Neural Networks, 27, 272-282.

[16] Al-Latief, S. T. A., Yussof, S., Ahmad, A., Khadim, S. M., & Alkhayyat, A. (2025). WAR Strategy Algorithm-based Hybrid Optimization for Accurate and Rapid Speech Recognition. Iraqi Journal for Computer Science and Mathematics, 6(1), 13.

[17] Shadiev, R., & Liu, J. (2023). Review of research on applications of speech recognition technology to assist language learning. ReCALL, 35(1), 74-88.

[18] Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). ‘Look, I can speak correctly’: learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. Computer Assisted Language Learning, 37(5-6), 1335-1363.

[19] Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2023). Exploring the effects of automatic speech recognition technology on oral accuracy and fluency in a flipped classroom. Journal of Computer Assisted Learning, 39(1), 125-140.

[20] Kim, S., & Jung, H. (2018). A study on the utilization of speech recognition technology in foreign language learning applications-focusing on English and French speech. Journal of Digital Contents Society, 19(4), 621-630.

[21] Luan, X. (2023, February). Development and Application of Spoken French Corpus Based on AI Speech Recognition. In 2023 IEEE 6th Eurasian Conference on Educational Innovation (ECEI) (pp. 115-118). IEEE.

[22] Papin, K., & Cardoso, W. (2025). Assessing the Pedagogical Potential of Google Translate's Speech Capabilities: Focus on French Pronunciation. calico journal, 42(1), 1-24.

[23] Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. Foreign Language Annals, 51(3), 617-637.