Research On Hydrodynamic Simulation And Speech Recognition Method Based On PLDA Model

Wu Lixian[1*], Lai Weiwei[2], Bu li[1], Lin Yujie[1], Wu Guangcai[2], and Zheng Yinglong[2]

[1] Foshan Power Supply Bureau of Guangdong Power Grid Co., Ltd., Foshan 528000, Guangdong, China
[2] Guangdong Electric Power Information Technology Co., Ltd., Guangzhou 510062, Guangdong, China
[*] Corresponding author. E-mail: wulixian016@outlook.com

Received: Mar. 25, 2026; Accepted: Apr. 28, 2026

Speech recognition is one of the important technologies of biological information recognition, which can extract the corresponding characteristics through language recognition, so as to recognize the generated speech. The International Institute of Technology (National Institute of Standards and Technology) evaluated speech recognition technology and found that the Probabilistic Linear Discriminant (PLDA), Analysis) model was goodHowever, in practice, speech recognition is easily affected by external factors such as environmental noise and human voice, leading to differences between registered and test speech and making data processing difficult. These challenges limit its application. To address issues of timelength variation and limited training samples, this study applies and refines a PLDA-based probabilistic correction model by adjusting its distribution using speech data. Finally, the PLDA parameters obtained by the training are taken as the test value, and the hydrodynamic simulation software is used to improve the performance of speech recognition. Hydrodynamic simulation enhances the PLDA model's robustness by addressing speech duration and cross-domain variability, leading to improved recognition accuracy. Experiments demonstrate notable gains in both EER and DCF metrics. It is found that the speech recognition method based on PLDA model can effectively improve the recognition function after hydrodynamic simulation, solve the problems of time length mismatch and limited training samples, and provide a theoretical basis for the application of PLDA model in speech recognition.

Keywords: PLDA model; speech recognition; hydrodynamic simulation; cross-domain migration

http://dx.doi.org/10.6180/jase.202609_32.057

1. Introduction

With the rapid development of the society and the continuous replacement of The Times, China has entered the information age. Identity verification plays an important role in modern times, but traditional ways like passwords and barcode systems are often employed; yet, they are prone to being stolen or misused, particularly for password-based approaches [1, 2]. In contrast, biometric authentication relies on unique physiological or behavioral traits like voice, fingerprints, or irises for enhanced accuracy and safety. Speech recognition stands out due to its accessibility and affordability. The human body has unique biological characteristics that are difficult to duplicate, making biometric keys secure and convenient as they cannot be easily forgotten or stolen [3]. Speech recognition is a biometric identification technology, similar to fingerprint and iris recognition, used for identity verification. Compared with other biometric methods, it is more convenient and cost-effective, as speech can be captured via mobile devices with low hardware requirements, while feature extraction and model training can be performed on a computer [4].

Speech recognition emerged in early-mid 20th century focusing on voiceprint analysis [5]. Later, research shifted to acoustic feature extraction and linear prediction methods [6–8]. Mid-to-late 20th century introduced sound dis-

crimination via Hidden Markov Model (HMM) [9]. Subsequently, Gaussian Mixture Model (GMM) and Universal Background Model (UBM) were effectively applied to model speech distributions [10]. These approaches have several limitations when dealing with variability among various speakers and transmission media. In order to solve this problem, Joint Factor Analysis (JFA) was proposed to account for the different types of variability. Building on this, the i-vector method provides an efficient low-dimensional representation of speech features. Moreover, the PLDA technique models variability between and within speakers. Several speech recognition technologies have emerged [11]. Joint Factor Analysis (JFA) models speech characteristics across subspaces and is widely used. Inspired by JFA, the ivector framework was proposed [12], offering lower-dimensional representations. I-vectors are often combined with LDA and WCCN. Later, PLDA was introduced, showing strong robustness in speaker recognition [13]. Several recent papers have emphasized important strides made in the field of automatic speech recognition for a variety of low-resource languages. This shows that current ASR systems can easily adapt to different acoustic environments [14]. Sri Harsha Grandhi proposed a hyper-heuristic-based stutter speech recognition system using Mel spectrogram, PLP features, and deep learning with WFST decoding, improving robustness and accuracy in irregular speech conditions [15].

This study improves speech recognition by addressing duration variation and limited data. A probabilistic PLDA correction reduces mismatches, while cross-domain migration with hydrodynamic simulation enhances robustness for speaker verification and identification.

Speech recognition includes two stages: training and recognition. During the training stage, the features and representations of speech should be extracted and stored in the database. During the recognition phase, the parameters present in the test speech were lateral and the results were judged, as shown in Figure 1.

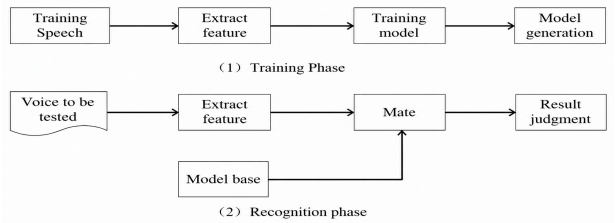

Fig. 1. Speech recognition framework diagram

Figure 1- Speech recognition framework diagram - illustrates the stages of speech recognition, including feature extraction, training, and testing; shows how speech signals are processed and mapped to PLDA parameters for recognition.

In feature extraction, speech features are derived from raw audio signals, including short-time inverse channel feature parameters analyzed over short durations. Speech is converted from analog to digital via sampling and quantization [14]. Common sampling rates include 8000 Hz, 12500 Hz, and 16000 Hz with linear quantization. Preprocessing includes pre-emphasis, framing, voice activity detection.

During pre-emphasis, high-frequency energy loss is compensated by applying a high-pass filter to enhance spectral balance, smooth the spectrum, and reduce noise, typically using 6 dB/oct (20 dB/dec), as shown in Equation 1.

$$G(s) = k\frac{sT}{1+sT} \tag{1}$$

For the software, the Z transfer functions are shown in Eq. 2.

$$H(z) = 1 - az^{-1} \tag{2}$$

A is indicated as the pre-aggravation coefficient. In the above equations, $x(n)$ represents the input speech signal, $y(n)$ denotes the output signal after pre-emphasis filtering, and n is the discrete time index.

Speech signals are segmented into frames before feature extraction to ensure smoothness, typically remaining stable within 30 ms . Window processing [16] is applied using rectangular and Hamming windows, as shown in formulas 3 and 4.

$$\omega(n) = \begin{cases} 1, 0 \le n \le N-1 \\ 0, \text{ other} \end{cases} \tag{3}$$

$$\omega_H(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 1, \text{ other} \end{cases} \tag{4}$$

Voice Activity Detection (VAD) identifies speech and non-speech regions in speech signals to improve recognition accuracy, reduce noise impact, and lower computational load [17]. Methods include energybased, model-based, sliding window, and multi-signal approaches. Energy-based uses signal energy, model-based uses trained speech/non-speech models, and sliding window applies thresholded fixed windows respectively.

Feature extraction of the speech signal is an important part of speech technology. Prior to this, raw speech signals are standardized through normalization and noise removal. Signal normalization ensures consistent amplitude levels,

while noise is reduced using filtering and voice activity detection (VAD). At present, the feature parameters existing in speech technology can be predicted linearly.

Mel-Frequency Cepstral Coefficients (MFCC) model human auditory perception of sound frequency. Humans are less sensitive to high frequencies and more to low frequencies. Below 1000 Hz , perception is linear; above 1000 Hz , it becomes logarithmic. The mapping relationship is shown in Equation 5 in the corresponding signal processing model framework.

$$F_{MEL} = 2595 * \lg\left(1 + \frac{F_{HZ}}{700}\right) \tag{5}$$

In equation 5, FHZ represents the current frequency, FMEL represents the perceived frequency, and the log-growth relationship is shown in Figure 2.

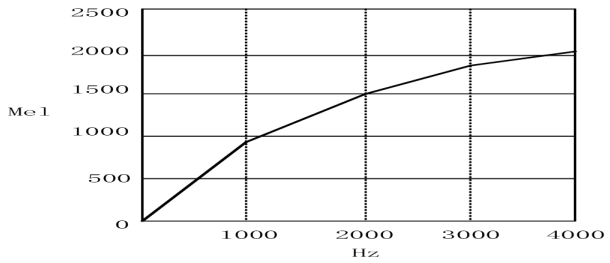

Fig. 2. Relationship between linear frequency and Mel frequency

Human auditory perception shows frequency-dependent sensitivity, with speech clarity most affected in the $200 - 5000$ Hz range [18]. Low-frequency sounds can mask higher frequencies, and critical bandwidth is narrower at low frequencies than high frequencies. Therefore, triangular Mel filter banks are used to match perception. MFCC extraction is illustrated in Figure 3.

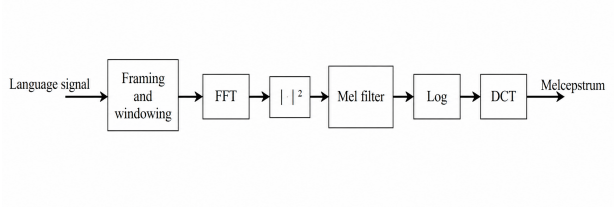

Fig. 3. M Flow chart of FCC parameter acquisition

Figure 3 - Flow chart of MFCC parameter acquisition. Shows the steps from speech signal preprocessing to frequency filtering, energy computation, and cosine transformation; highlights how MFCC features are extracted for use in the PLDA model.

The MFCC feature extraction process can be described in a structured sequence as follows:
(1) Preprocessing: The voice signal goes through pre-emphasis, framing, and windowing processes.
(2) Spectral analysis: Short-time Fourier transform is used for spectral analysis.
(3) Mel-filtering: The filtered output from the spectrum goes through a bank of filters based on the Melscale for mimicking human hearing.
(4) Log energy: The energies obtained from the Mel-filters are put on a logarithmic scale.
(5) DCT transformation: Log energies are transformed to get MFCCs.

Speech signal is preprocessed into short-time spectrum. Spectrum squared yields energy, and M-band Mel filter bank computes band energies producing output power spectrum x(k). Logarithm applied, then cosine transform obtains MFCC coefficients as shown in Formula 6 respectively for extraction.

$$C_n = \sum_{K=1}^{M} \log x(k) \cos\left[\pi(k - 0.5)\frac{n}{M}\right], n = 1, 2, L \tag{6}$$

In Equation 6, M is the number of filters, x(k) is the filter output, and L denotes the filter order. Channel effects introduce variations and distortion in speech signals, though channels remain stable over short periods [19]. Inverted spectrum parameters are relatively invariant, and convolutional noise can be reduced using a high-pass filter [20]. Mitigation methods include CVN, Feature Warping, Feature Mapping, and CMS [21]. This study uses CMS and CVN; CMS is given in Equation 7.

$$C'_d = C_d(t) - \frac{1}{N}\sum_{i=1}^{N} C_d(i); d = 1, 2, \cdots, D \tag{7}$$

In Equation 7, Cd(t) represents the feature component, D represents the feature dimension, and N represents the total number of frames.

CNV is mainly the error after the cheap channel of spectral features, see Equation 8.

$$C'_d(t) = \frac{C_d(t)}{\sigma_d}; d = 1, 2, \cdots, D \tag{8}$$

σ_d In formula 8, the standard variance in d is indicated. Generally speaking, CMS and CVN are used to improve the robustness of the system.
Consider $\omega_{i,j}$ as the j speech i -vector vectors of the i -source, as shown in formula 9.

$$\omega_{i,j} = \mu + \phi\beta_i + Gw_{i,j} + \varepsilon_{i,j} \tag{9}$$

In the formula, u represents the mean of the i -vector, ϕ represents the matrix, which can also be used as a subspace for speech data processing, and G represents the channel subspace, β I represents speech related implicit variables, wi, j represents the speaker's implicit variables, and $\varepsilon_{i,j}$ represents residuals. In addition, the PLDA model will also merge Gwi, j , and $\varepsilon_{i,j}$ as a noise term, as shown in formula 10 .

$$\omega_{i,j} = \mu + \phi\beta_i + \varepsilon_{ij} \tag{10}$$

The basis for establishing a PLDA model is to input data and perform EM calculations on the parameters.

The EM algorithm uses an iterative approach to estimate the parameters of the model:
(1) Expectation step (E-step): calculates the expectation of the latent variables based on the current parameter estimates.
(2) Maximization step (M-step): updates the parameters by maximizing the likelihood function with the results from the E-step.

If there are N speakers in the development set, speaker i has Mi speech segments and i -vector representation. EM algorithm estimates PLDA parameters via maximum likelihood. E-step computes hidden variable expectations iteratively. When J = 1 + M and 1 is identity matrix, formulas 11 and 12 are derived.

$$E\left(\beta_i\right) = J^{-1}\phi\sum^{-1}\phi \tag{11}$$

$$E\left(\beta_i\beta_i^T\right) = E\left(\beta_i\right)E\left(\beta_i\right)^T + J^{-1} \tag{12}$$

If two i-vector ($1\omega, \omega 2$) are determined, they are used as test speech and trained speech respectively, and the two hypothesis models are obtained from the logarithm for specific calculation, see Equation 13.

$$\text{score } = \log P\left(\omega_1, \omega_2 \mid H_0\right) - \log P\left(\omega_1, \omega_2 \mid H_1\right) \tag{13}$$

Where H0 is the same source and H1 is different sources.

Due to the influence of speech duration, if $\mathcal{E}_{i,j}$ follows a normal distribution, then

$$\varepsilon_{i,j} - N\left(0, \left(\frac{L_{ij}}{a}\right)^{-\lambda}\sum\right) \tag{14}$$

Where L ij is the j speech duration of the i speaker and a is the tuning parameter.

If there are N speakers in the development set and Mi speech segments in the i speech, the i-vector can be represented by η_{ij}, and the posterior probability can be expressed as formula 15 .

$$P\left(\eta_{ij} \mid \beta_i\right) = N\left(\eta_{ij}\middle|\, \phi\beta_{i,}\left(\frac{L_{ij}}{a}\right)^{-\lambda}\sum\right) \tag{15}$$

After introducing the intermediate variable K , the delay probability can be calculated by the Bayesian rule, see Equation 16.

$$P\left(\beta_i \mid F_i\right) = N\left(\beta_i \mid K^{-1}\phi\sum^{-1}M_i'F_i, K^{-1}\right) \tag{16}$$

The speech feature distribution is modeled as $x \sim \mathcal{N}(\mu, \Sigma)$, where μ and Σ are the mean and covariance of *i*-vector features. To reduce duration and channel mismatch, hydrodynamic simulation applies $x' = Ax + e$, where A is the transformation matrix and $e \sim \mathcal{N}\left(0, \Sigma_h\right)$ is residual noise. The resulting distribution is $x' \sim \mathcal{N}\left(A\mu, A\Sigma A^T + \Sigma_h\right)$, used in the PLDA framework for improved robustness.

The probabilistic correction model adjusts the i-vector distribution using a transformation matrix and residual noise to handle variability, and the corrected representation is then integrated into the PLDA framework for improved speaker modeling.

The overall process of the PLDA probabilistic correction algorithm can be outlined as follows:
Stage 1: Input of speech signal and preprocessing (Pre-emphasis, Framing, and VAD)
Stage 2: MFCC feature extraction and i-vector representation, where MFCC features are used to generate i-vector representations that capture speaker characteristics and are then processed by the probabilistic correction model within the PLDA framework."
Stage 4: Probabilistic correction by taking into account duration and channel variation
Stage 5: Modify i-vector distribution with the use of a correction model
Stage 6: Verification through obtained modified PLDA parameters
Stage 7: Recognition results output and evaluation of performance (EER, DCF)

An ablation study is conducted to evaluate the effectiveness of the probability-corrected PLDA method. The performance of the baseline i -vector model, standard PLDA, and the proposed PLDA correction model is compared. Results show progressive improvement across models, with the proposed method achieving the best EER and DCF values, confirming the effectiveness of the probability correction mechanism. Statistical significance testing is performed to validate the improvements in EER and minDCF. A paired t -test is applied between the baseline models and the proposed PLDA correction model. The results indicate that

Table 1. Experimental Parameters and Configuration

Parameter	Description	Value/Setting
Feature Type	Acoustic feature extraction	MFCC
i-vector Dimension	Speaker representation size	400
PLDA Type	Backend classifier	Standard PLDA
Transformation Matrix (A)	Correction mapping	Learned matrix
Residual Noise (Σh)	Variability modeling	Gaussian noise
Optimization Method	Model tuning approach	Maximum likelihood
Evaluation Metrics	Performance measures	EER, minDCF
Dataset	Evaluation corpora	NIST SRE 2008/2010

Table 2. NIST SRE08 EER values for comparison

duration	I-vector	I-vector+PLDA	The PLDA probabilistic correction model
Full length	10.9%	7.9%	7.1%
20s	14.9%	11.7%	11.3%
10s	17.8%	13.6%	13.4%

Table 3. NIST SRE08 DCF08 values for comparison

duration	I-vector	I-vector+PLDA	The PLDA probabilistic correction model
Full length	0.0496	0.0362	0.0360
20s	0.0614	0.0526	0.0528
10s	0.0693	0.0648	0.0644

the performance improvements are statistically significant ($p < 0.05$), confirming that the observed gains are not due to random variation.

Computational complexity is considered to evaluate the scalability of the proposed approach. The training phase has higher computational complexity due to iterative PLDA parameter estimation, whereas the inference phase has relatively low complexity as it involves only similarity computation.

A sensitivity analysis is conducted on the tuning parameter α to evaluate its effect on EER and minDCF. The results indicate that performance remains stable within a suitable range of α, confirming the robustness of the PLDA correction model.

Hydrodynamic simulation is used to evaluate the PLDA correction model under varying speech durations, noise conditions, and real-world environmental variability to assess its robustness and performance stability.

In order to verify the set probability correction model, it was simulated by hydrodynamics, and the software NIST-SRE 08 and NISTSRE10 were selected for performance testing. With the help of 32dimensional UBM-GMM trained in MFCC and 512, i -vector dimension is 400 , and the dimension of PLDA speech factor βi is 120 . To reduce the influence of external factors, different GMM mixture orders and data variations are considered. Cross-domain migration is used to bridge the distributional differences between NIST-SRE 2008 and NIST-SRE 2010 datasets, enabling consistent PLDA performance across heterogeneous evaluation conditions. Dataset preparation includes a structured recording setup, participant variation, and channel diversity across devices and environments to ensure completeness of the experimental data and improve robustness under real-world speech conditions.

The NIST SRE08/SRE10 datasets include SwitchBoard female data, UBM, T-matrix, and PLDA training models. NIST04/05 provides 1077 UBM, 6063 T-matrix, 47027 PLDA data, 1140 short2 training models, 1674 short3 tests, and 1800 iterations with 21580 recognition tests. NISTSRE10 core includes 11370 UBM, 20348 T-matrix and PLDA data.

NIST SRE08 and SRE10 were truncated to 10s and 20s for evaluation. UBM, T-matrix, and PLDA were used with fixed parameters. Three systems (i -vector baseline, PLDA +i -vector, and probability- corrected PLDA) were tested, with α tuned $0.7 - 0.9$ using mean duration. A sensitivity analysis of parameters α and λ shows their impact on system performance across different speech durations and testing conditions. The results indicate that proper tuning of these parameters is essential, as variations lead to noticeable changes in EER and DCF performance.

Using Equal Error Rate (EER) and Minimum Decision Cost Function (minDCF), i-vector, PLDA, and probability correction systems were evaluated under varying durations in NIST-SRE08. EER increases with duration Table 2, while DCF08 increases with shorter duration Table 3, showing improved robustness of probability correction.

Table 4. NIST SRE10 EER values for comparison

duration	I-vector	I-vector+PLDA	The PLDA probabilistic correction model
Full length	8.2%	4.1%	3.7%
20s	12.7%	7.0%	6.5%
10s	17.7%	10.4%	9.9%

Table 5. NIST SRE10 DCF08 values for comparison

duration	I-vector	I-vector+PLDA	The PLDA probabilistic correction model
Full length	0.0310	0.0362	0.0178
20s	0.0507	0.0320	0.0311
10s	0.0610	0.0438	0.0464

In NIST SRE10, EER values, DCF 08 values increased with the length decrease, as shown in Tables 4 and 5.

NIST-SRE08 and NIST-SRE10 show similar trends, indicating PLDA probability correction improves identification performance. Under full-time NIST-SRE10 conditions, the probability-corrected model achieves significantly better EER and DCF08 values compared to the other two recognition systems, demonstrating superior robustness and accuracy overall. Compared with existing i-vector and standard PLDA approaches, the proposed method shows improved performance in terms of lower EER and DCF across different speech durations. This highlights its uniqueness in effectively handling temporal and cross-channel variability through duration-aware correction. These improvements in EER and DCF indicate a reduction in both false acceptance and false rejection rates, enhancing overall system accuracy. In practical applications, this leads to more reliable and robust speaker verification performance under varying speech durations and channel conditions.

2. Materials and methods

PLDA-based speech recognition relies on development and training data, but limited datasets reduce performance and robustness. Existing resources from LDC, OGI, and Chinese corpora (dialogue, scanning, tourism) support database creation, yet channel variability and single-channel recordings limit effectiveness. This study proposes a cross-channel voiceprint recognition database to improve diversity. The dataset includes information, number strings (6 repetitions, 18 items), short passages (20 texts), and 10 dialogue topics with $\sim$ 1-minute responses per dialogue. This structured design of the speech corpus ensures phonetic balance and linguistic diversity by incorporating varied speech contents such as numerical strings, short passages, and spontaneous dialogues, thereby improving the generalization ability and reliability of the model.

The recording setup includes Skype, PowerGramo, and Steinberg Cubase 5 software, along with microphones, headsets, in-ear headsets, mobile phones, and recording pen. Five voice input methods are used, including Skype and mobile inputs, as shown in Figure 4.

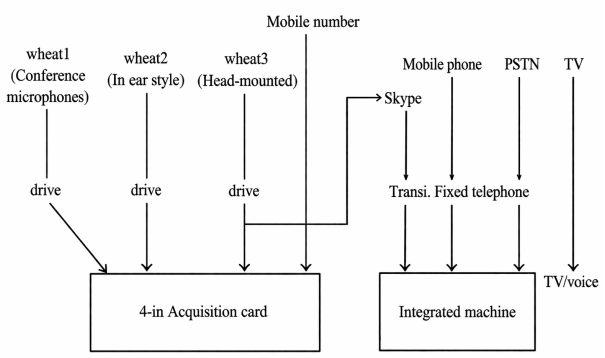

Fig. 4. Enter the system design drawing

Figure 4 presents the system configuration of cross-channel speech data acquisition involving the devices used (microphones, headphones, in-the-ear headphones, and mobile phones), modes of data input (Skype, recording pen), and software involved (Steinberg Cubase 5, PowerGramo). It emphasizes the importance of collecting speech data from 100 people systematically in order to train and test PLDA models.

The cross-channel voiceprint database includes speech data from 100 speakers recorded across multiple input channels such as microphones, mobile phones, headsets, and Skype. Balanced sampling was maintained, and channel variability was assessed based on differences in noise levels and transmission characteristics, ensuring sufficient diversity for reliable experimental evaluation.

A total of 100 subjects aged 20-24 was recorded in a quiet room speaking Mandarin. Participants provided name, age, and throat discomfort status before recording. Post-recording, manual inspection was conducted to prevent data entry errors and ensure dataset accuracy [22, 23].

Speaker variability factors such as age and speaking rate were considered during data acquisition to reduce dataset bias and improve reliability. The participants were selected within a consistent age group, and all subjects were instructed to maintain a natural and uniform speaking pace.

3. Results and discussion

There are two databases, A and B. Database A provides large training data for PLDA, while B has limited data. Direct transfer from A to B causes mismatch. A cross-domain migration technique is proposed to adapt the PLDA model from A to B effectively. The PLDA model parameters in database A are set to ($\varphi_B, \sum_B$), and the PLDA model parameters in target database B are set to ($\varphi_B, \sum_B$). There is a certain relationship between matrices φA and φB in the relevant subspaces A. $\sum$B is the variance matrix, which can be fused on the model, as shown in formulas 17 and 18.

$$\phi_r = a\phi_A + (1-a)\phi_B \tag{17}$$

$$\sum_r = a\sum_A + (1-a)\sum_B \tag{18}$$

Among them, ϕ_r represents the subspace after model fusion, $\sum$r represents variance, and a is the adjustment parameter.

The study uses real network data as database A and cross-channel speech recognition data as database B . Database A contains 22,000 voice samples from microphone and telephone channels. A 512-dimension UBM, 400-dimension T matrix, and PLDA model were trained and evaluated in Table 6.

Table 6. Minimum DCF value comparison

φ_A And $\sum$ A test	φ B and $\sum$ B test	φ r and $\sum$ r test
0.2675	0.0844	0.0610

After data comparison, the PLDA model φr and $\sum$ The r test identified it even better, while the φ B and $\sum$B is not very good. The reason for this analysis may be due to the limited data samples in the B database, and the PLDA model obtained after training is difficult to show speech features. φ_A And $\sum$A The test effect is the worst, and the reason is mainly because of the difference between databases, which directly leads to the low data matching degree.

Figure 5 shows loud-region duration correlations with speaker traits. It has moderate positive correlation with age (Spearman $= 0.53$), strong negative with speaking rate (Spearman $= -0.76$), and very strong with utterance

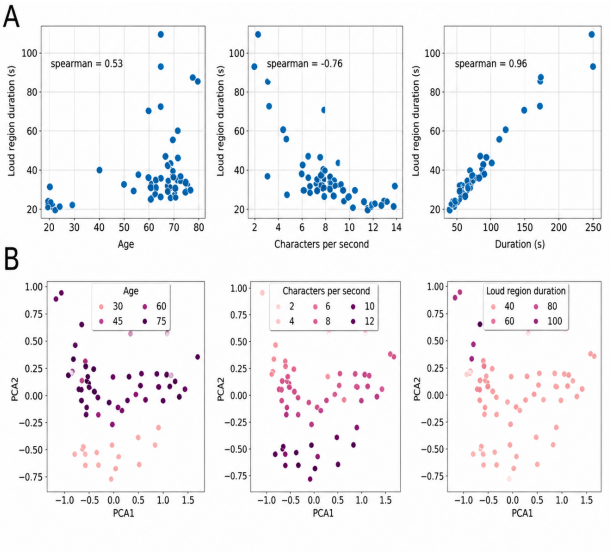

Fig. 5. Speech temporal feature correlation and PCA visualization.

duration (Spearman $= 0.96$). PCA of i -vector/PLDA embeddings shows speaker space shifts, supporting duration-aware PLDA correction. Temporal speech attributes like speaking rate and duration affect i-vector/PLDA embedding distribution, reducing clarity and increasing intra-speaker variability.

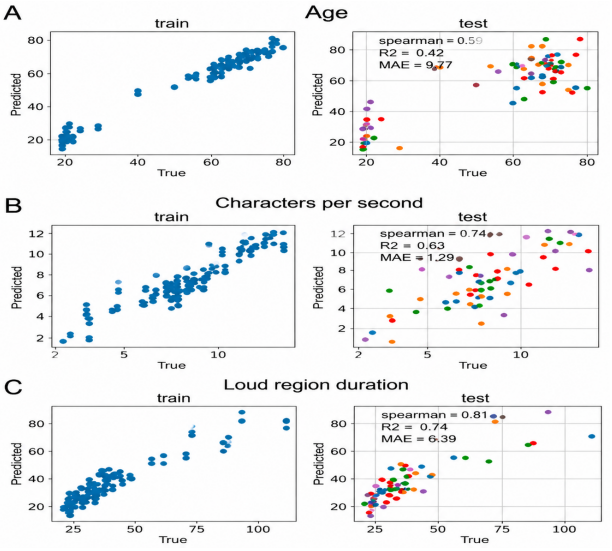

Fig. 6. Prediction performance of temporal speech attributes using i-vector/PLDA representations.

Figure 6 presents prediction performance for age (A), speaking rate (B), and loud-region duration (C). Training data shows strong linear alignment between true and predicted values using i-vector/PLDA embeddings. In test set,

age (Spearman $= 0.59, R^2 = 0.71$), speaking rate ($0.74, 0.63$), and loud-region duration $(0.81, 0.74)$ indicate increasing performance respectively.

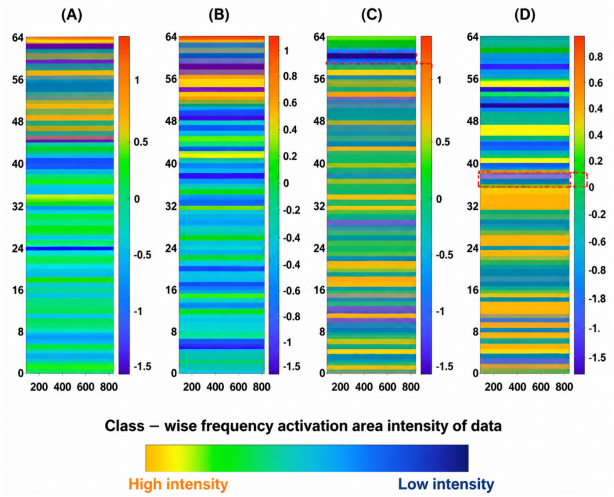

Fig. 7. Class-wise frequency activation intensity patterns across speaker groups.

Figure 7 shows stable mid-to-high frequency activation with lower-frequency variability. Panels (A)-(B) align, while (C)-(D) reveal increased high-frequency variability and PLDA divergence under mismatched conditions.

4. Conclusion

Voice recognition uses voiceprint features for identity authentication with simple operation, low cost, wide applicability, and high security. However, systems are sensitive to external factors causing duration inconsistency and channel challenges. This study proposes a hydrodynamic PLDA-based approach to improve recognition performance, summarized in two main conclusions.

(1) For the problem of time length inconsistency, this study proposed the PLDA probability correction model, and trained it in detail, and found that the problem of time length mismatch can be effectively solved, so as to improve the speech recognition performance.

(2) For the high difficulty in channel acquisition, this study proposes constructing a cross-channel voice library and using a cross-domain PLDA recognition method to improve recognition performance.

Acknowledgment

Declarations

Data Availability
Not applicable

Conflicts of Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution
Wu Lixian: Conceptualization, methodology, investigation, writing—original draft.
Lai Weiwei: Data curation, software, validation.
Buli: Formal analysis, visualization.
Lin Yujie: Software, simulation, resources.
Wu Guangcai: Review and editing, project administration.
Zheng Yinglong: Supervision, critical revisions.
All authors reviewed and approved the final manuscript.

Ethical Approval
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to Participate
Not applicable.

Consent to Publication
All authors consent to the publication of this manuscript.

Competing Interests
The authors declare no competing interests.

References

[1] H. Lauren, (2023) *"Incorporating Automatic Speech Recognition Methods into the Transcription of Police-Suspect Interviews: Factors Affecting Automatic Performance"* **Frontiers in Communication** ***8***: DOI: 10.3389/fcomm.2023.1165233.

[2] Z. Dong, Q. Ding, W. Zhai, and M. Zhou, (2023) *"A Speech Recognition Method Based on Domain-Specific Datasets and Confidence Decision Networks"* **Sensors** ***23***(13): DOI: 10.3390/s23136036.

[3] H. Wang, Z. Li, D. Song, X. He, and M. Khan, (2022) *"Applying Machine Learning and Automatic Speech Recognition for Intelligent Evaluation of Coal Failure Probability under Uniaxial Compression"* **Minerals** ***12***(12): 1548. DOI: 10.3390/min12121548.

[4] J. Wu, Y. Zhang, L. Xie, Y. Yan, X. Zhang, S. Liu, X. An, E. Yin, and D. Ming, (2022) *"A Novel Silent Speech Recognition Approach Based on Parallel Inception Convolutional Neural Network and Mel Frequency Spectral Coefficient"* **Frontiers in Neurorobotics** ***16***: 971446. DOI: 10.3389/fnbot.2022.971446.

[5] M. Amini, D. Matrouf, J. F. Bonastre, S. Dowerah, R. Serizel, and D. Jouvet. "Learning Noise Robust ResNet-Based Speaker Embedding for Speaker Recognition". In: *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. 2022. DOI: 10.21437/Odyssey.2022-6.

[6] E. H. Alkhammash, M. Hadjouni, and A. M. Elshewey, (2022) *"A Hybrid Ensemble Stacking Model for Gender Voice Recognition Approach"* **Electronics *11***(11): 1750. DOI: 10.3390/electronics11111750.

[7] J. Heeren, T. Nuesse, M. Latzel, I. Holube, V. Hohmann, K. C. Wagener, and M. Schulte, (2022) *"The Concurrent OLSA Test: A Method for Speech Recognition in Multi-talker Situations at Fixed SNR"* **Trends in Hearing *26***: 23312165221108257. DOI: 10.1177/23312165221108257.

[8] K. Zhao and D. Wang, (2021) *"Research on Speech Recognition Method in Multi-Layer Perceptual Network Environment"* **International Journal of Circuits, Systems and Signal Processing *15***: 996–1004. DOI: 10.46300/9106.2021.15.107.

[9] J. Guglani and A. Mishra, (2020) *"Automatic Speech Recognition System with Pitch-Dependent Features for Punjabi Language on KALDI Toolkit"* **Applied Acoustics *167***: DOI: 10.1016/j.apacoust.2020.107386.

[10] Y. H. Tu, J. Du, and C. H. Lee, (2019) *"Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing *27***(12): 2080–2091. DOI: 10.1109/TASLP.2019.2940662.

[11] H. B. Prasetio, H. Tamura, and K. Tanno, (2019) *"Generalized Discriminant Methods for Improved X-Vector Back-end Based Stress Speech Recognition"* **IEEJ Transactions on Electronics, Information and Systems *139***(11): 1341–1347. DOI: 10.1541/ieejeiss.139.1341.

[12] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani. "Learning the Speech Front-End with Raw Waveform CLDNNs". In: *Interspeech*. 2015, 1–5. DOI: 10.21437/Interspeech.2015-1.

[13] S. Cumani, O. Plchot, and P. Laface. "Probabilistic Linear Discriminant Analysis of i-Vector Posterior Distributions". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, 7644–7648. DOI: 10.1109/ICASSP.2013.6639150.

[14] J. Guglani and A. Mishra, (2020) *"Automatic Speech Recognition System with Pitch-Dependent Features for Punjabi Language on KALDI Toolkit"* **Applied Acoustics *167***: DOI: 10.1016/j.apacoust.2020.107386.

[15] S. H. Grandhi and M. M. Kamruzzaman, (2024) *"Automatic Stutter Speech Recognition and Classification Using Hyper-Heuristic Search Algorithm"* **International Journal of Automation and Smart Technology *14***(1): DOI: 10.5875/dm761m36.

[16] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. Wong, and M. J. Gales, (2019) *"Exploiting Future Word Contexts in Neural Network Language Models for Speech Recognition"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing *27***(9): 1444–1454. DOI: 10.1109/TASLP.2019.2922048.

[17] Y. Zhang, P. Zhang, and Y. Yan, (2019) *"Language Model Score Regularization for Speech Recognition"* **Chinese Journal of Electronics *28***(3): 604–609. DOI: 10.1049/cje.2019.03.015.

[18] C. Li, H. Ge, and G. Chen. "A Robust Speech Feature Extraction Method Based on Nonlinear Power Transform Gammachirp Filter". CN109256127B. 2021.

[19] N. Brümmer, A. Swart, L. Mošner, A. Silnova, O. Plchot, T. Stafylakis, and L. Burget, (2022) *"Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for Length-Normalized Embeddings"* **arXiv preprint**: DOI: 10.48550/arXiv.2203.14893.

[20] B. Priyanka and S. I. A., (2018) *"A Probabilistic Feature-Based SVM Model for Hindi/English Speech Recognition"* **International Journal of Engineering & Technology *7***(2.8): 271. DOI: 10.14419/ijet.v7i2.8.10423.

[21] V. R. Raman and G. J. Vysotsky. "Methods and Apparatus for Generating and Using Garbage Models for Speaker Dependent Speech Recognition Purposes". US5842165A. 1998.

[22] Y. Jeong, (2013) *"Speaker Adaptation Using Probabilistic Linear Discriminant Analysis for Continuous Speech Recognition"* **Electronics Letters *49***(25): 1641–1643. DOI: 10.1049/el.2013.2223.

[23] C. Zhang, B. H. Roh, and G. Shan. "Poster: Dynamic Clustered Federated Framework for Multi-Domain Network Anomaly Detection". In: *Companion of the 19th International Conference on Emerging Networking Experiments and Technologies*. 2023, 71–72. DOI: 10.1145/3624354.3630086.