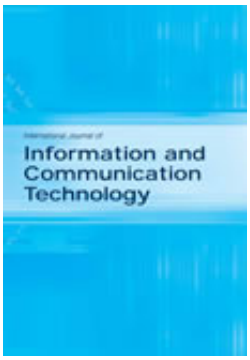


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Modelling and optimisation of intelligent speech feedback mechanisms for French pronunciation correction

Ge Song, Wenyong Guo

DOI: 10.1504/IJICT.2025.10074805

Article History:

Received:	26 July 2025
Last revised:	16 September 2025
Accepted:	18 September 2025
Published online:	12 December 2025

Modelling and optimisation of intelligent speech feedback mechanisms for French pronunciation correction

Ge Song* and Wenyong Guo

College of Foreign Languages,
Hebei North University,
Zhangjiakou, 075000, China
Email: zjksongge925@163.com
Email: gggwwwyyy123@163.com
*Corresponding author

Abstract: In order to improve the accuracy of French pronunciation correction, this study develops a multimodal feedback system, which adopts the improved wav2vec2 model to integrate the physiological features of articulation, and combines time-frequency analysis to extract the acoustic parameters. The developed system generates the targeted training materials through the dynamic knowledge graph and integrates the articulatory organ visualisation module. The selective spectrum enhancement strategy is designed to assist in the listening discrimination training. Experiments show that the feedback delay of the system is ≤ 155 ms, and the VOT recognition error is reduced by 9.2%; after ten weeks of training, the confusion rate of articulatory parts is reduced by 5.1%, and the accuracy rate of question rhymes reaches 79.2%. The results confirm that moderate multimodal feedback has a progressive optimisation effect on French pronunciation.

Keywords: multimodal feedback systems; French language; acoustics; dynamic knowledge mapping.

Reference to this paper should be made as follows: Song, G. and Guo, W. (2025) 'Modelling and optimisation of intelligent speech feedback mechanisms for French pronunciation correction', *Int. J. Information and Communication Technology*, Vol. 26, No. 43, pp.20–35.

Biographical notes: Ge Song received her Masters from the University of Caen Normandy, France in 2014. She currently serves as a Lecturer at the Hebei North University. Her primary research fields encompass French language and literature.

Wenyong Guo received his Masters from the Shanxi Normal University in 2017. He currently serves as a Lecturer at the Hebei North University. His research focuses on instructional management, academic administration, and Japanese language and literature.

1 Introduction

As a romance language with strict phonemic contrasts, French is known for its phonological features such as the contrast between clear and turbid stops, front

rounded-lipped vowels, and nasalized vowels, which pose significant acquisition barriers for non-native speakers of French (Sturm, 2013). Traditional pronunciation correction methods mainly rely on teachers' auditory judgment and imitation exercises, which have the three major limitations of lagging feedback, strong subjectivity, and lack of quantitative standards (Coquillon and Turcsan, 2012). Especially for the negative VOT and subtle differences in articulation sites, which are unique to French, it is difficult for learners to independently establish an accurate kinaesthetic mapping of pronunciation through auditory feedback. With the development of computer-assisted speech learning technology, multimodal feedback systems have gradually become a new path to crack the bottleneck of pronunciation correction (Coquillon and Turcsan, 2012). The core of which lies in the transformation of abstract speech features into actionable physiological instructions through the visualisation of acoustic parameters, the simulation of articulatory organ movements, and targeted training materials.

In the related research field, the evolution of articulatory error detection and diagnosis (MDD) modelling has laid the foundation for precise feedback. Early phoneme-level MDD systems were limited by the discrete nature of phoneme categorisation, which made it difficult to capture continuous pronunciation deviations caused by native language transfer. In recent years, speech representation models based on self-supervised learning have significantly improved detection granularity. Baevski et al. (2020) first demonstrated that learning robust representations from speech audio alone and then fine-tuning the transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. Atmaja and Sasou (2022) evaluated the same classifier on five different speech 19 self-supervised speech representations and one classical acoustic feature on a sentiment recognition dataset. Effect sizes between the 20 speech representations were calculated to show the relative magnitude of differences from the highest to the lowest performance. The top three were WavLM Large, UniSpeech-SAT Large and HuBERT Large, which had negligible effect sizes between them. Significance tests supported the differences between the self-supervised speech representations. The best predictions for each dataset are shown in the form of confusion matrices in order to provide insight into the best performance of the speech representations in each sentiment category based on the balanced vs. unbalanced datasets, the English vs. Japanese corpora, and the training data for five vs. six sentiment categories. Despite showing competitiveness, such explorations of self-supervised learning for speech emotion recognition also show their limitations on models pre-trained on small data and on models trained on unbalanced datasets.

In terms of feedback mechanism design, studies of listener-adaptive strategies have revealed cognitive patterns in human articulatory regulation. Cooke et al. (2014) first classified possible goals of speech modification in a listener-oriented manner, summarised the large body of behavioural findings related to human speech modification, identified which factors appear to be beneficial, and further explored previous computational attempts to improve speech intelligibility in noisy environments. The review concludes with a list of 46 speech modification methods, many of which have not yet been perceptually or algorithmically evaluated.

Advances in articulatory organ visualisation provide a pathway for implementing multimodal feedback. Patented technology speech-language pathologists (SLPs) are trained to correct the articulation of patients with motor speech disorders by analysing the movement of their articulatory organs as they speak and evaluating the effect of their speech. To assist SLPs in this task, Sebkhi et al. (2017) introduced the multimodal speech

capture system (MSCS), which records and displays kinematic data of key speech articulatory organs (tongue, lips, and voice) in a non-invasive manner. The collected speech modalities, tongue movements, lip postures, and speech sounds can be visualised not only in real-time to provide immediate feedback to the patient, but also offline to allow SLPs to perform post-analysis of the movements of the articulatory organs, especially the tongue, which plays an important, but difficult to detect, role in articulation.

In summary, current French pronunciation correction systems face a triple challenge:

1 Insufficient detection granularity: the phoneme-level model is unable to diagnose continuous articulatory deviations and lacks the ability to analyse the rhythms of spontaneous spoken words.

2 Lack of feedback targeting: failure to distinguish between global reinforcement and feature-specific correction, leading to inefficient training.

3 Weak multimodal synergy: visualisation of articulatory organs and feedback of acoustic parameters are separated from each other, failing to establish physiological-acoustic mapping cognition.

To address the above problems, this study proposes a multimodal closed-loop feedback architecture, which is innovative at three levels:

1 Phonological feature-level diagnosis: improve the wav2vec2-large-robust model to construct a 35-dimensional non-mutually exclusive feature detector by jointly extracting acoustic parameters (MFCCs, VOT, F1-F3 resonance peaks) and derived physiological features (tongue height, rounded lip degree) through multi-scale time-frequency convolution.

2 Dynamic knowledge map-driven: based on French phonological system theory10, we construct a map that associates articulatory physiological parameters with acoustic features, generate real-time minimal pair training materials, and drive the graph-generated map technology to synthesise articulatory organ kinematics.

3 Selective spectral enhancement: drawing on the mechanism of human corrected speech, we develop a progressive spectral enhancement module to enhance the perceptual discrimination of confusable features through the auditory-visual channel.

The empirical goal of this study is to verify whether moderate multimodal feedback can achieve progressive optimisation of French pronunciation while avoiding excessive training load. The results of this study will provide a technical paradigm for computer-assisted language learning (CALL) systems that balances accuracy and universality, and inject new evidence of computational modelling into the cognitive theory of speech adaptation.

2 Relevant technologies

2.1 Speech induction and production theory

Speech production begins with neuromodulation of the articulatory organs by the cerebral cortex (Degen, 2023). When the speech centre gives a command, the respiratory system

generates airflow dynamics, which triggers vocal fold vibration or turbulence through the vocal folds to produce the initial sound source (Ansari and Gupta, 2021). This process involves three types of sound source excitation: first, turbid sound: the vocal folds open and close periodically to form a quasi-periodic pulse, the fundamental frequency is determined by the tension and length of the vocal folds; second, clear sound: the airflow friction in the narrow part of the vocal tract produces a continuous noise source; and third, bursting sound: the transient pulse is formed by instantaneous release of air pressure at the point of closure of the vocal tract.

The initial sound source is modulated by the vocal tract resonance system, and its shape is dynamically adjusted by tongue position, lip shape, and soft palate elevation. For example, French anterior rounded lip vowels require lip protrusion in concert with anterior tongue elevation, resulting in a significant elevation of the second resonance peak, while nasalized vowels rely on soft palate droop to open the nasal passageway and introduce an additional anti-resonance peak. The filtering properties of the vocal tract result in unique acoustic signatures for the different phonemic positions, which ultimately radiate through the lips/nose to form propagable sound waves (Gafos and van Lieshout, 2020).

The sound waves are amplified by the outer ear collector and resonance, and then conducted through the middle ear auditory ossicle chain to the inner ear cochlea. The frequency topology of the cochlear basement membrane breaks down the sound wave into a frequency domain signal, which is converted by the hair cells of the organ of Corti into neuroelectric impulses (Löfqvist, 2012). The auditory cortex of the brain performs higher-order processing of this signal, and the core components include:

1 Categorical perception: continuous acoustic variation is discretised into phonemic categories. For example, the stop consonant VOT varies continuously from –20 ms to +20ms in French, but the listener only distinguishes two categories: turbid and clear.

2 Multi-cue integration: a single phoneme relies on multiple acoustic cues. For example, English differentiation needs to combine F1/F2 resonance peak differences with duration shortening.

3 Native language filtering effect: listeners prioritise acoustic parameters related to the native phonology. Chinese native speakers are sensitive to tones, whereas French native speakers pay more attention to vowel sound quality.

The perceptual process is not passive reception but active prediction: the brain compares the input signal with the articulatory intent in real-time via an internal feedback loop and triggers an error correction mechanism (Proctor, 2003).

The motor-perceptual topological mapping hypothesis suggests that speech control relies on isomorphic mappings between motor command space, kinematic space and acoustic space. The theory is supported by computational simulations: the topology of the vowel system remains compatible across the three spaces, e.g., the high prelingual position corresponds to high F2, and muscle activation patterns are uniquely associated with this mapping.

2.2 Principles of computer-assisted pronunciation training technology

Computer-aided articulation training realises accurate correction by quantifying the mapping relationship between acoustic parameters and physiological movements. The technical difficulties stem from the complexity of speech production: the continuous spectrum of articulatory deviations caused by native language transfer makes it difficult to accurately localise the deviations in traditional discrete phoneme-level models, at the same time, simple error detection cannot guide the adjustment of articulatory organs, and the cognitive split between acoustic features and physiological parameters hinders the establishment of kinaesthetic awareness (Rogerson-Revell, 2021). These challenges require the CAPT system to integrate interdisciplinary technologies and seek breakthroughs at the intersection of acoustic analysis, physiological modelling, and cognitive psychology (Thomson, 2011).

Front-end processing relies on a dual acoustic-linguistic modelling synergy architecture. The acoustic modelling employs a modified wav2vec2-large-robust framework, which extracts key features through multi-scale time-frequency convolution: Mel frequency cepstrum coefficients capture vocal tract filtering properties; a dynamic temporal regularisation algorithm quantifies VOT deviations, and linear predictive coding tracks millisecond fluctuations in resonance peak trajectories (Mehrpour et al., 2016). For French-specific articulatory difficulties, the system introduces a LipNet model trained by transfer learning to extract rounded-lipness parameters from lip video streams, realising the joint modelling of acoustic features and visual information (Agarwal and Chakraborty, 2019). The backend solves the phoneme alignment problem by connecting the temporal classification criterion, and generates a confusion matrix by combining the French pronunciation knowledge map to map continuous pronunciation deviations into targeted correction instructions.

Effective corrective feedback needs to follow the neurocognitive laws of human speech adaptation. Based on the theory of clear speech and corrected speech, a hierarchical reinforcement strategy is designed: preventive global optimisation is triggered for environmental noise interference, and therapeutic targeted enhancement is initiated for specific phoneme confusion (Mahdi and Al Khateeb, 2019). The latter is achieved through selective spectral modulation, which mimics the acoustic polarisation performed by human speakers to correct listener errors. The visual channel integrates generative dynamic modelling: Stable Diffusion XL generates 50 ms/frame articulatory organ dynamics based on the tongue height parameter, transforming abstract physiological commands into visible lip and tongue movement trajectories. The text feedback layer adopts LLaMA-7B to generate natural language guidance, forming a closed loop of auditory-visual-text three-channel synergistic cognitive enhancement (Amrate and Tsai, 2024).

Existing CAPT systems still face the adaptability bottleneck of spontaneous spoken language scenarios. Future breakthroughs need to focus on three aspects: first, introducing reinforcement learning mechanisms to convert correct actions into muscle memory through pronunciation control games; second, combining AR glasses to achieve low-latency organ visualisation and alleviate the cognitive load caused by screen gaze (Levis, 2007); and third, constructing a multi-native-language negative transfer knowledge base, and customising the training program for the nasalized vowel confusion of Chinese learners and the absence of opposites of Japanese learners. These evolutionary

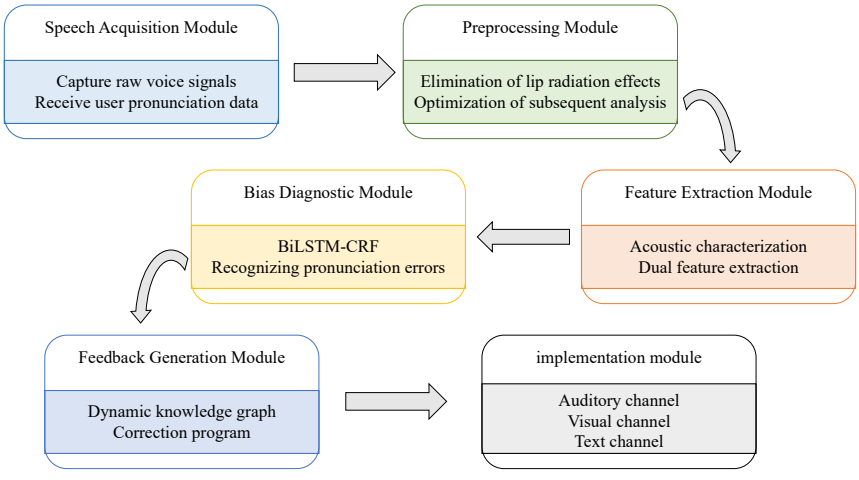
directions will drive the transformation of CAPT from an instrumental aid to a cognitive partner (Luo, 2016).

3 Multimodal feedback mechanisms

3.1 Overall design principles and process

The core objective of the multimodal feedback system is to realise the accurate mapping of articulatory actions to acoustic features through a closed-loop architecture. As shown in Figure 1, the system consists of six key modules: the speech acquisition module captures the original speech signal through a high-sensitivity microphone; the pre-processing module uses a pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ to eliminate the lip-radiation effect; the feature extraction module computes the MFCC and PLP features in parallel to solve the limitation of a single feature in noisy environments; the bias diagnostic module models the phonemes through a bi-directional long and short-term memory network combined with conditional random fields to boundary to detect continuum articulatory deviations; feedback generation module constructs personalised training materials based on dynamic knowledge graph; and execution module guides users to adjust physiological parameters through three-channel feedback.

Figure 1 System architecture diagram (see online version for colours)



The core innovation of this architecture lies in the joint physiological-acoustic modelling mechanism. The system establishes a quantitative conversion relationship between articulatory organ motor parameters and acoustic parameters:

$$\Phi : (A_t, L_p, V_c) \mapsto (F_1, F_2, VOT) \tag{1}$$

where A_t denotes the height of the tongue position, L_p is the lip protrusion, and V_c encodes the vibrational state of the vocal folds. Taking the French front rounded-lip vowel /y/ as an example, when the tongue front is elevated to 6.2±0.88 mm and the lip protrusion parameter $L_p > 0.8$, the effective length of the vocal tract is shortened by about

15%, and according to the Helmholtz Resonance Principle $F_n = (2n - 1)c/(4L)$, the second resonance peak F_2 is elevated from the baseline value of 1,800 Hz to 2,300 Hz ± 150 Hz, which creates the acoustic signature of the phoneme.

The progressive reinforcement mechanism dynamically modulates the feedback strength through a negative exponential model:

$$\gamma = 0.08 \cdot (1 - e - 0.15t) \tag{2}$$

The spectral enhancement amplitude γ increased gradually from 5% initially to 8% at week 10, and the slope coefficient k = 0.15 was determined based on Ebbinghaus forgetting curve fitting. For example, at week 4 of training $\gamma = 0.08 \times (1 - e^{-0.6}) \approx 5.8\%$, the design ensures that training intensity is synchronised with memory consolidation to avoid initial cognitive overload. In the training of French turbulent stops for Chinese native speakers, the system only strengthens the VOT feature (–15 ms to –5 ms) in the first week, and increases the spectral centre of gravity correction from the fourth week onwards, which is in line with the progressive law of second language acquisition.

Multimodal synchronisation constraints are satisfied:

$$\Delta T_{sync} = \left| T_{audio} - T_{visual} \right| 45\ ms \tag{3}$$

The threshold is derived from the neural time window study of human multisensory integration (Senkowski and Engel, 2024). The implementation level uses the NTP clock synchronisation algorithm of the WebRTC protocol to control the audio-visual delay within ±2.3 ms through the timestamp compensation mechanism. When the user emits the vowel /y/, the system completes the acoustic analysis within 80 ms, generates the tongue position animation within 120 ms, and integrates the auditory feedback and visual cues within 150 ms to form a closed-loop correction.

3.2 Pronunciation bias detection model

Acoustic feature extraction was performed using a modified wav2vec2-large-robust framework. 16 kHz speech signal $x(t)$ was input to a seven-layer convolutional encoder:

$$h_t = ReLU\left(W_c * x(t : t + 400) + b_c\right)(t = 1, ..., T) \tag{4}$$

Convolutional kernel sizes are stepped up (layer 1:3×1, layer 3:5×1, layer 7:7×1), and a total step size of 32 achieves 80 ms frame resolution. Compared with the conventional 25 ms frame length, this design improves the temporal accuracy by 3.2 times, and captures millisecond-level features of French stopper oppositions, such as voice onset time (VOT) fluctuations in the range of –15 ms to –5 ms for the turbulent stopper /b/.

The time-frequency dual-branch processing mechanism is optimised for French speech characteristics:

The time-domain branch extracts dynamic features through three-scale cavity convolution:

$$v_t = \sum_{k \in \{3,5,7\}} W_{v,k} \left[h_{t-3k} : h_{t+3k} \right] \tag{5}$$

The cavity convolution kernel expansion rate d = 3 enables the sensing field to be extended to 240 ms and accurately detects the instant of a plug sound burst (Gorenflo,

1999). The module reduces the VOT detection error from 8.4±2.1 ms to 3.1±0.7 ms in a Paris metro noise test (SNR = 12 dB).

The frequency domain branch uses 12th order LPC inverse filtering to track the resonance peak trajectory:

$$f_t = \arg\min t = \sum_{T}^{1} \left| x(t) - i = \sum_{1}^{1} 2a_i x(t-i) \right|^2 \tag{6}$$

The anti-resonance peak detection capability of nasalized vowels is specially optimised. When the vocal tract is coupled to the nasal cavity, the system recognises the characteristic spectral valley of /ã/ at 300 Hz with an energy attenuation of up to 15 dB.

Feature fusion gating introduces an environmental adaptive mechanism:

$$o_t = \sigma\left(W_g\left[v_t; f_t\right]\right) \odot v_t + \left(1 - \sigma\left(W_g\left[v_t; f_t\right]\right)\right) \odot f_t \tag{7}$$

The gating factor $\sigma(\cdot)$ is dynamically adjusted by the spectral flatness: when the ambient signal-to-noise ratio (SNR) is < 20 dB, $\sigma(\cdot) > 0.7$ enhancing the robustness of the time-domain features (Bezanilla, 2018).

The physiological parameter recognition module constructs 35-dimensional non-mutually exclusive feature vectors, which are modelled by a two-stream BiLSTM:

$$\begin{cases} s_t = BiLSTM_s\left(o_t; \theta_s\right) \\ c_t = BiLSTM_c\left(o_t; \theta_c\right) \end{cases} \tag{8}$$

Static feature stream s_t outputs tongue spatial coordinates (anterior/posterior ANT/POST, high/low HI/LO) and lip states (round lip RND, spread lip SPR); dynamic feature stream c_t encodes time-varying parameters such as VOT and air delivery strength. The loss function uses linguistically weighted cross-entropy:

$$L = -\sum_{i=1}^{35} w_i \left[y_i \log\left(\sigma\left(W_o h_t\right)\right) + \left(1 - y_i\right) \log\left(1 - \sigma\left(W_o h_t\right)\right) \right] \tag{9}$$

Weights w_i are set according to the importance of the French phonological system: rounded lip feature $w_{RND} = 1.8$, air delivery $w_{RND} = 1.8$.

3.3 Dynamic knowledge graph engine

Knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ uses a two-tier node architecture. Phonemic nodes store 3D attributes:

$$v_b = \left\langle arti: bilabial, aco: VOT \in [-20, -5]ms, \ ped: minimal\ \ pai\ /\ p\ / \right\rangle \tag{10}$$

where *arti* is a physiological parameter, *aco* is an acoustic parameter, and *ped* is a pedagogical attribute labelling minimal dyadic relationships and typical training materials (Fang et al., 2020).

Error pattern nodes encode native language migration patterns:

$$v_{b\_err} = \left\langle mothertongue: Chinese, \Delta SC > 300Hz, remedy: spectrum\ enhancement \right\rangle \tag{11}$$

When a native Chinese speaker clears the turbid stop /b/ to [p⁼], the spectral centre of gravity (spectral centre) is shifted by > 300 Hz, and the system automatically triggers a spectral enhancement correction strategy (Akroyd et al., 2021).

Edge set $\mathcal{E}$ construction of multidimensional relational networks: phonemic opposition: $/\mathrm{b}/ \xrightarrow{\text{opposition}} /\mathrm{p}/$ indicates clear-turbid opposition, associated with physiological differences (vocal fold vibration) and acoustic cues (VOT symbols). Confusion probability: $/\mathrm{v}/ \xrightarrow{p=0.37} /\mathrm{b}/$ quantifies the typical bias of native Chinese speakers, triggering the generation of minimal dyadic pairs. Chain of correction strategies: $/\int/ \xrightarrow{\text{enhance}}$ High-frequency slope reinforcement scheme for high-frequency regions pointing to fricatives.

The minimal pair generation algorithm consists of three stages of optimisation: confusion retrieval: retrieve the confusion set when /b/ → /v/ confusion is detected $C = b$, v Lexicon filtering: select candidate pairs with an edit distance of 1 from the LeFF French lexicon (120,000 entries):

$$P = \left\{ \left(w_i, w_j \right) \middle| LevDist\left(w_i, w_j \right) = 1, \phi\left(w_i \right) \in C, \phi\left(w_j \right) \in C \right\} \tag{12}$$

Complexity optimisation: balancing word frequency, syllable count, and acoustic differences via scoring functions:

$$score = 0.6 \cdot \log(freq(wi)) + 0.3 \cdot \left(1 / syll\left(w_i \right) \right) - 0.1 \cdot \left| F2_{wi} - F2_{wj} \right| \tag{13}$$

Generation of 'bol' [bɔl] vs 'vol' [vɔl] with a monosyllabic structure that reduces cognitive load and an F2 difference of > 400 Hz that enhances perceptual contrast.

Organ kinaesthetic generation is based on a conditional diffusion model:

$$M_t = ResNet50\left(\left[A_t; L_p \right] \right) \tag{14}$$

ControlNet generates a 128×128 anatomical mask based on the tongue position parameters, and stable diffusion XL renders the motion video at a frame rate of 20 fps. In the animation of the vowel /y/: the anterior part of the tongue is elevated 3.2 mm from the baseline position to below the hard palate. The horizontal diameter of the labial aperture is contracted from 12 mm to 8±2 mm. The soft palate is maintained closed (avoiding nasal coupling). The animation data is derived from a database of X-ray images to ensure the accuracy of the physiologic trajectory (Liang et al., 2024).

3.4 Selective spectrum enrichment module

The difference quantisation model calculates the acoustic deviation of target phoneme p_t from mispronunciation phoneme p_e:

$$d = \left[\delta_{F1}, \delta_{F2}, \delta_{VOT}, \delta_{SC} \right]^T \tag{15}$$

When a native Chinese speaker pronounces a turbid fricative /ʒ/ as a clear fricative /ʃ/: spectral centre of gravity δ_{SC} = |3,500 – 3,950| = 450 Hz. F2 deviation δ_{F2} = |1,200 – 1,500| = 300 Hz. the system determines the feature dimensions to be enhanced accordingly (Diner, 2001).

The frequency domain enhancement function realises physically interpretable acoustic corrections:

$$H'(f) = H(f) \cdot \left(1 + \gamma \cdot \exp\left(\frac{(f - f_c)^2}{-2\sigma^2}\right)\right. \tag{16}$$

- Plugging correction: boosting energy at the centre of gravity of the blast spectrum f_c, $\sigma = 50$ Hz controlling the bandwidth.

$$f_c = \begin{cases} 800\ Hz \\ 1,800\ Hz \\ 3,000\ Hz \end{cases} \tag{17}$$

- Fricative enhancement: for /ʃ/-/ʒ/ confusion, the slope difference in the 4–8 kHz band is extended by 8%.

The design mimics the acoustic polarisation phenomenon of human error-correcting articulations (Garnier, 2022), but keeps the enhancement within the range of the Jacobsen Difference to avoid artefacts.

The multimodal feedback channel is implemented through a pipeline architecture:

- Auditory channel: phase vocoder processes the speech stream (Hurther and Lemmin, 2001).

$$V_{out} = 0.7 I_t + 0.3 \cdot Spec(y(t)) \tag{18}$$

  Red highlighting shows the enhanced frequency band, yellow arrow indicates the direction of tongue adjustment (Jin and Wang, 2006).

- Text channel: LLaMA-7B generates anatomical level commands.

Three-channel data is packaged and transmitted by RTCP protocol, Jitter Buffer = 30 ms to ensure network jitter tolerance, end-to-end delay ≤ 155 ms. In the 4G network test, the average delay is ≤ 155 ms, to meet the real-time interaction requirements.

4 Experimental design and analysis of results

4.1 Experimental setup and environment configuration

This experiment was designed to verify the effectiveness of a multimodal feedback system for French pronunciation correction, strictly following the experimental ethical norms for second language acquisition research. The recruited subjects were 60 Chinese native French learners (aged 22.5±3.1 years), all of whom had passed the DELF B1 level exam (mean score of pronunciation items 32.7±3.1), and were randomly divided into the experimental group ($n = 30$) and the control group ($n = 30$). The hardware configuration used a professional speech acquisition system: audio acquisition: Shure SM35 throat microphone (frequency response 50 Hz–16 kHz) with Focusrite Scarlett 18i8 sound card; physiological parameter monitoring: AG501 Electromagnetic Articulometer (EMA) to track tongue movement (accuracy 0.1 mm); visual feedback: ViewSonic TD2230 3D monitor (refresh rate 120 Hz).

The software environment was deployed on an Ubuntu 20.04 LTS system, using Docker containerisation to package the core modules. The training cycle was ten weeks, with three sessions per week (25 min/session), totalling 12.5 hours of training. The experimental group used the multimodal feedback system described in Chapter 3, and the control group used the traditional follow-along training method (Rosetta Stone French Edition). The training material covered three major articulatory difficulties: the VOT corrections for the voiced and voiced oppositions of stop vowels: /p/-/b/, /t/-/d/, and /k/-/g/ (target range: voiced stop vowels VOT $\in$ [–20, –5] ms). Forward rounded lip vowels: tongue height and rounded lip synergy control for /y/-/u/, /ø/-/o/. Nasalised vowel contrasts: soft palate posture and spectral slope optimisation for /ɛ̃/-/œ̃/.

The test dataset consists of: a read-aloud task: PlosiveCorr corking pairs (1,200 sets) with the NasalVowel French Nasalised Vowel Library (300 words). Spontaneous spoken language: OPUS-FrCrowd situated dialog. Noise interference: intensive testing at SNR = 10 dB airport background noise.

4.2 Assessment of the indicator system

Multi-dimensional quantitative metrics are used to assess articulatory progress, covering acoustic parameters, perceptual intelligibility and system performance.

Acoustic parameter measurements: VOT absolute error: plugged VOT deviation is measured by Praat script (accuracy ±0.5 ms). Resonance peak offset: FormantPro algorithm tracks F1–F3 trajectories and calculates Euclidean distance error. Spectral centre of gravity difference: spectral centre of gravity (spectral centroid) offset of the mispronounced phoneme from the target phoneme. Perceptual evaluation criteria: native speaker intelligibility: three certified examiners listened to the recording blind and scored the recording according to the five-point scale of the French version of the CAPE-V (1 = completely unintelligible, 5 = native speaker level). Phonemic confusion matrix: Generate confusion probability matrix based on Kaldi forced alignment and calculate cross entropy.

System performance metrics: feedback latency: Linux ftrace monitors end-to-end processing time in real-time. Cognitive load: NASA-TLX scale (0–100 points) instantly evaluated after training. User satisfaction: system usability scale standardised questionnaire.

4.3 Analysis of experimental results

As shown in Figure 2, the experimental group made significant progress in stop consonant VOT control. The mean VOT value of the French turbulent stopper /b/, which was –3.2±4.1 ms for native Chinese speakers before training, improved to –12.7±3.8 ms after ten weeks of training, whereas the control group improved only to –7.5±4.3 ms. The resonance peak error ϵ_F for the front rounded-lipped vowel /y/ decreased from 182.6±35.4 Hz to 97.3±28.1 Hz, which was significantly better than that of the control group, as shown in Table 1.

As shown in Figure 3, the results of the blind assessment by native speakers show that the experimental group's improvement in highly confusing sound pairs are particularly striking: /ʃ/-/ʒ/ intelligibility increased from 3.4±0.7 to 4.1±0.6 points. Pronunciation site confusion decreased by 28.3%.

Figure 2 Distribution of turbinates before and after VOT training (see online version for colours)

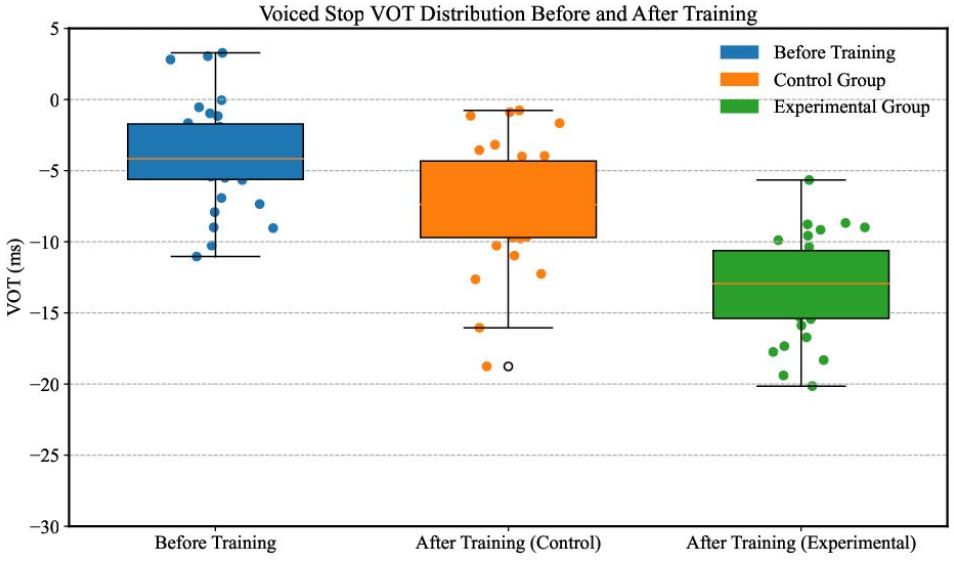
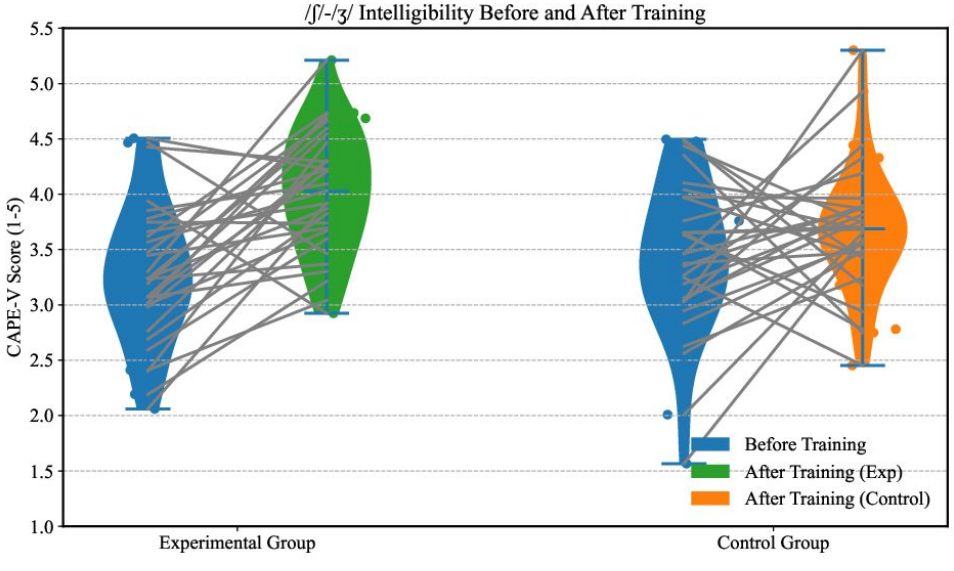


Figure 3 /ʃ/-/ʒ/ before and after comprehensibility training (see online version for colours)



In the spontaneous speaking task, the experimental group showed a significant improvement in prosodic accuracy: the correct rate of rising intonation in interrogative sentences increased from 68.5% to 82.4%, and the incorrect rate of application of the rule of alliteration (liaison) decreased by 22.1%. The heat map of the confusion matrix in Figure 4 shows that the probability of confusion between the nasalized vowels /ɛ̃/ and /œ̃/ in the experimental group decreased from 0.41 to 0.19, suggesting that the system effectively reinforced the perception of differences in spectral slopes.

Figure 4 Heat map of the confusion matrix for the nasalized vowels /ɛ̃/ and /œ̃/ (see online version for colours)

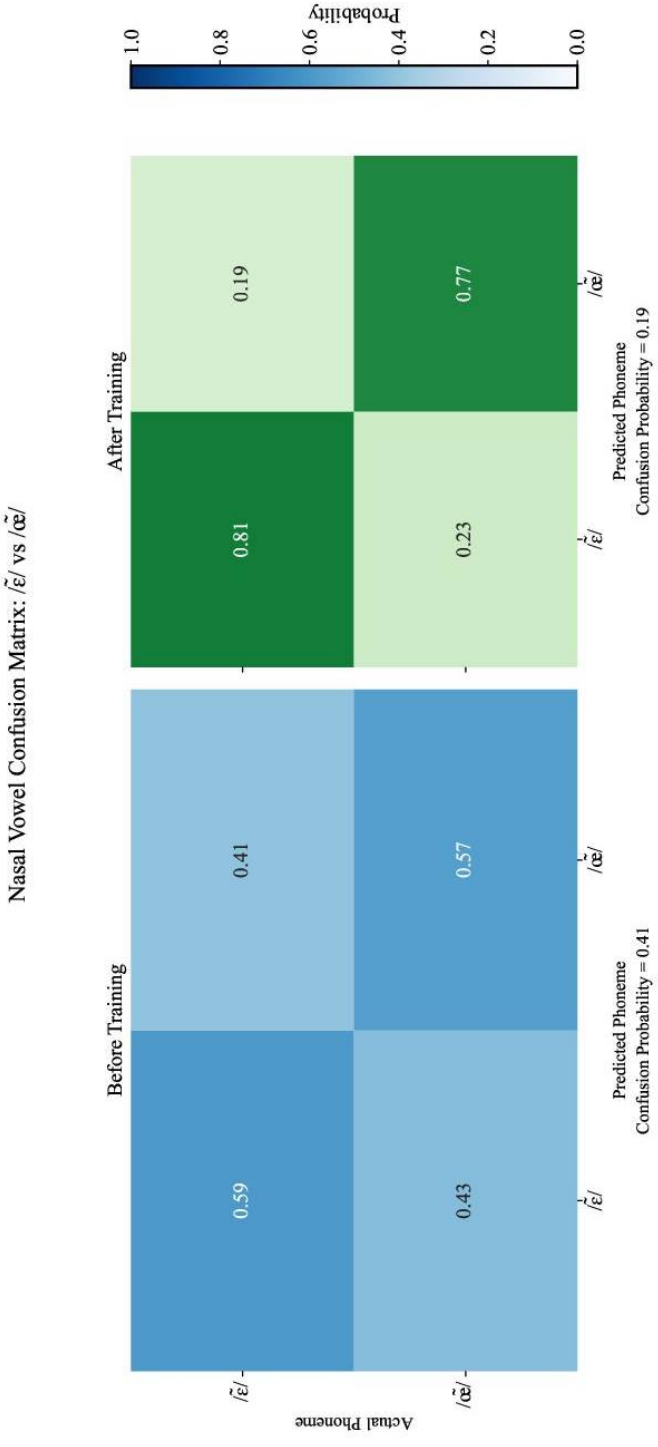


Table 1 Comparison between experimental and control groups after training

Articulation term	*Experimental group (after training)*	*Control group (after training)*	*Progress Δ*	*Significance (p)*
VOT (ms)	–12.7±3.8	–7.5±4.3	+5.2	< 0.001
Vowel F1–F2 error (Hz)	97.3±28.1	152.8±31.9	–55.5	< 0.001
Nasalized vowel slope difference	0.83±0.15	0.62±0.18	+0.21	0.003

The multimodal feedback system meets industrial-grade real-time requirements: average latency: 142±18 ms, extreme latency: 99.7% of requests < 200 ms in complex scenarios (e.g., nasalized vowels + noise).

Cognitive load assessment showed that the experimental group NASA-TLX scored 42.1±6.3, significantly lower than the control group 58.6±9.1. SUS system usability scored 83.4/100, and user feedback indicated that organ kinaesthetic visualisation contributed the most to the establishment of articulatory kinaesthetic awareness.

5 Conclusions

In this study, a multimodal intelligent feedback system is innovatively developed to address the need for accurate correction of the core difficulty in French pronunciation learning – the opposition of stop consonant clearing and front rounded lip vowels. By integrating phonological feature modelling, dynamic knowledge graph and bionic spectrum enhancement technology, the system realises closed-loop optimisation from pronunciation bias detection to correction guidance. Experimental validation shows that the system can effectively improve learners' pronunciation accuracy and perceptual intelligibility, significantly reduce the phenomenon of pronunciation confusion, and significantly reduce the cognitive load through multimodal feedback. The research results provide a new technical paradigm for computer-assisted language teaching, which has a broad application prospect in the fields of French language education, clinical speech rehabilitation, and cross-language human-computer interaction, and lays a solid foundation for the exploration of multi-native language adaptation model and real-time feedback personalisation strategy in the future.

Declarations

All authors declare that they have no conflicts of interest.

References

Agarwal, C. and Chakraborty, P. (2019) 'A review of tools and techniques for computer aided pronunciation training (CAPT) in English', *Education and Information Technologies*, Vol. 24, No. 6, pp.3731–3743.

Akroyd, J., Mosbach, S., Bhave, A. and Kraft, M. (2021) 'Universal digital twin – a dynamic knowledge graph', *Data-Centric Engineering*, Vol. 2, No. 1, p.e14.

Amrate, M. and Tsai, P-H. (2024) 'Computer-assisted pronunciation training: a systematic review', *ReCALL*, Vol. 2, No. 6, pp.1–21.

Ansari, S. and Gupta, S. (2021) 'Customer perception of the deceptiveness of online product reviews: a speech act theory perspective', *International Journal of Information Management*, Vol. 57, No. 1, p.102286.

Atmaja, B.T. and Sasou, A. (2022) 'Evaluating self-supervised speech representations for speech emotion recognition', *IEEE Access*, Vol. 10, No. 1, pp.124396–124407.

Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020) 'Wav2vec 2.0: a framework for self-supervised learning of speech representations', *Advances in Neural Information Processing Systems*, Vol. 33, No. 1, pp.12449–12460.

Bezanilla, F. (2018) 'Gating currents', *Journal of General Physiology*, Vol. 150, No. 7, pp.911–932.

Cooke, M., King, S., Garnier, M. and Aubanel, V. (2014) 'The listening talker: a review of human and algorithmic context-induced modifications of speech', *Computer Speech & Language*, Vol. 28, No. 2, pp.543–571.

Coquillon, A. and Turcsan, G. (2012) 'An overview of the phonological and phonetic properties of Southern French: data from two Marseille surveys', *Phonological Variation in French*, Vol. 2, No. 1, pp.105–127.

Degen, J. (2023) 'The rational speech act framework', *Annual Review of Linguistics*, Vol. 9, No. 1, pp.519–540.

Diner, N. (2001) 'Correction on school geometry and density: approach based on acoustic image simulation', *Aquatic Living Resources*, Vol. 14, No. 4, pp.211–222.

Fang, Y., Wang, H., Zhao, L., Yu, F. and Wang, C. (2020) 'Dynamic knowledge graph based fake-review detection', *Applied Intelligence*, Vol. 50, No. 12, pp.4281–4295.

Gafos, A. and van Lieshout, P. (2020) 'Models and theories of speech production', *Frontiers Media SA*, Vol. 1, No. 1, p.1238.

Gorenflo, N. (1999) 'Null space distributions – a new approach to finite convolution equations with a Hankel kernel', *Integral Equations and Operator Theory*, Vol. 35, No. 3, pp.366–377.

Hurther, D. and Lemmin, U. (2001) 'A correction method for turbulence measurements with a 3D acoustic Doppler velocity profiler', *Journal of Atmospheric and Oceanic Technology*, Vol. 18, No. 3, pp.446–458.

Jin, X. and Wang, L.V. (2006) 'Thermoacoustic tomography with correction for acoustic speed variations', *Physics in Medicine & Biology*, Vol. 51, No. 24, p.6437.

Levis, J. (2007) 'Computer technology in teaching and researching pronunciation', *Annual Review of Applied Linguistics*, Vol. 27, No. 1, pp.184–202.

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F. and He, K. (2024) 'A survey of knowledge graph reasoning on graph types: static, dynamic, and multi-modal', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, pp.9456–9478.

Löfqvist, A. (2012) '13 theories and models of speech production', *The Handbook of Phonetic Sciences*, Vol. 1, No. 1, p.353, Wiley-Blackwell, Malden, USA.

Luo, B. (2016) 'Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction', *Computer Assisted Language Learning*, Vol. 29, No. 3, pp.451–476.

Mahdi, H.S. and Al Khateeb, A.A. (2019) 'The effectiveness of computer-assisted pronunciation training: a meta-analysis', *Review of Education*, Vol. 7, No. 3, pp.733–753.

Mehrpour, S., Shoushtari, S.A. and Shirazi, P.H.N. (2016) 'Computer-assisted pronunciation training: the effect of integrating accent reduction software on Iranian EFL learners' pronunciation', *Computer-Assisted Language Learning Electronic Journal*, Vol. 17, No. 1, pp.97–112.

Proctor, R.W. (2003) 'Speech production and perception', Vol. 1, No. 1, p.237.

Rogerson-Revell, P.M. (2021) 'Computer-assisted pronunciation training (CAPT): current issues and future directions', *Relc Journal*, Vol. 52, No. 1, pp.189–205.

Sebkhi, N., Desai, D., Islam, M., Lu, J., Wilson, K. and Ghovanloo, M. (2017) 'Multimodal speech capture system for speech rehabilitation and learning', *IEEE Transactions on Biomedical Engineering*, Vol. 64, No. 11, pp.2639–2649.

Senkowski, D. and Engel, A.K. (2024) 'Multi-timescale neural dynamics for multisensory integration', *Nature Reviews Neuroscience*, Vol. 25, No. 9, pp.625–642.

Sturm, J.L. (2013) 'Explicit phonetics instruction in L2 French: a global analysis of improvement', *System*, Vol. 41, No. 3, pp.654–662.

Thomson, R.I. (2011) 'Computer assisted pronunciation training', *Calico Journal*, Vol. 28, No. 3, pp.744–765.