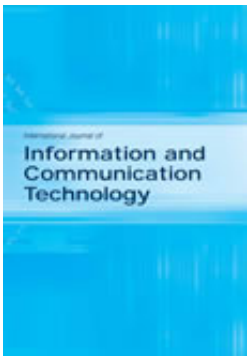
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Japanese pronunciation detection and corpus construction based on cross-modal attention

Xiaolu Liu

DOI: 10.1504/IJICT.2025.10074807

Article History:

Received:	25 September 2025
Last revised:	21 October 2025
Accepted:	25 October 2025
Published online:	12 December 2025

Japanese pronunciation detection and corpus construction based on cross-modal attention

Xiaolu Liu

Global Language Center,
Xi'an Eurasia University,
Xi'an, 710065, China
Email: lululaoshu@163.com

Abstract: To address Japanese pronunciation error detection, this paper proposes a fusion method based on cross-modal attention mechanisms and constructs a Japanese pronunciation corpus. The model integrates audio Mel-spectrogram and visual lip-motion features through attention mechanisms, effectively capturing fine-grained cross-modal interactions and enabling precise phoneme-level error recognition. Evaluated on both the public corpus from Saruwatari Lab, University of Tokyo and a self-built corpus, the proposed approach achieves an accuracy of 92.3%, which is 3.1% higher than the best baseline model. Moreover, it maintains a robust accuracy of 85.3% under a low signal-to-noise ratio of 5 db, representing a 6.6% improvement compared to other methods. This study provides an effective and noise-robust tool for multimodal speech learning with strong potential for educational applications. The released corpus contains 50 hours of multimodal data with detailed annotations, offering comprehensive support for Japanese language teaching and advanced speech technology development.

Keywords: cross-modal learning; pronunciation error detection; Japanese speech processing; attention mechanisms; corpus construction.

Reference to this paper should be made as follows: Liu, X. (2025) 'Japanese pronunciation detection and corpus construction based on cross-modal attention', *Int. J. Information and Communication Technology*, Vol. 26, No. 43, pp.61–77.

Biographical notes: Xiaolu Liu is a Lecturer in the Global Language Center at Xi'an Eurasia University, China. She obtained her Bachelor's in Japanese (2009) and a Master's in Japanese Language and Literature (2012) from Xi'an International Studies University, China. Her research interests include Japanese linguistics, Japanese speech processing, and corpus construction.

1 Introduction

As one of the world's major languages, the demand for Japanese phonetics learning has grown steadily in recent years (Aldossari et al., 2025), particularly highlighting its importance in cross-language communication and educational applications (Dailey, 2006). Accurate pronunciation is not only the foundation of effective communication but also a critical component in foreign language instruction. However, Yi-Ping and Allport (1995), Japanese's phonological system presents several unique challenges for non-native

learners (Cheng, 2022), such as sensitivity to moraic structure, voicing opposition, and the voicelessness of specific vowels (Ringen, 1999). Common errors include consonant confusion, mispronunciation of long/short vowel distinctions, and intonation deviations. If not corrected promptly (Zhang et al., 2010), these mistakes can significantly impair the intelligibility and naturalness of linguistic communication (Benot and Goff, 1998).

In the technological advancement of pronunciation error detection, traditional approaches primarily rely on automatic speech recognition systems to diagnose errors by comparing acoustic differences between learners' speech outputs and standard pronunciations (HOYT and Kenneth, 1987). While such methods have achieved some progress in languages like English (Sciarinigourianova, 2002), they exhibit a series of limitations when applied to Japanese. Specifically, approaches relying solely on audio signals struggle to effectively handle complex error patterns at the phoneme and suprasegmental levels (Octoplus, 2006). These include error masking caused by phonetic variation and co-articulation effects (Viswanathan et al., 2013), as well as reduced reliability in noisy environments. Furthermore, purely acoustic models inadequately reflect the physiological mechanisms of pronunciation, limiting their ability to explain errors and their applicability in teaching contexts (Diamond, 2013). The technical implementation of mispronunciation detection is indeed more challenging for Japanese relative to many other languages. This is primarily attributable to its Mora-timed rhythmic structure and the critical role of phonemic contrasts, such as those between geminate (double) and singleton consonants, and the distinction between long and short vowels, which are less prevalent or absent in many other languages and are particularly challenging for learners to master and for models to accurately assess.

In recent years, multimodal learning approaches have offered new insights for speech detection (Chung et al., 2016). By integrating visual information, particularly lip movement features, it is possible to capture physiological and acoustic correlations during speech production more comprehensively (Savariaux et al., 1995). Visual signals not only effectively supplement audio limitations in noisy environments but also provide crucial information about articulatory organ movements – such as lip shape (Ali et al., 2005), tongue position visibility, and jaw opening degree – which hold significant value for distinguishing specific phonemes. Although existing research has explored the potential of multimodal fusion in languages like English and Chinese (Cambria et al., 2013) systematic studies tailored to the phonetic characteristics of Japanese remain relatively scarce (Shigeno, 1986). This research direction is particularly important, especially considering the strong reliance on visual information for Japanese phonemes such as the labial consonant 'm' and the rounded vowel 'u'.

The broader context for this work is the rising prominence of cross-modal learning within contemporary artificial intelligence research. This paradigm, which focuses on integrating and aligning information from diverse sensory sources (e.g., audio and vision), has shown significant promise in enhancing model robustness and perceptual understanding, forming a key trend in developing more intelligent and adaptable AI systems. It is noteworthy that existing techniques still face critical challenges in multimodal representation learning (David-Pfeuty, 2006), particularly in achieving effective alignment and interaction between audio and visual signals (Arulanandam, 1994). Conventional fusion methods such as early fusion or late fusion often fail to fully exploit the fine-grained complementary relationships between modalities. The emergence of attention mechanisms offers a novel solution for dynamic cross-modal interaction. Through attention weight allocation, the model can adaptively focus on multimodal

feature segments most relevant to pronunciation errors, thereby enhancing detection accuracy and robustness (Hu, 2024). However, the application of this mechanism in Japanese pronunciation detection tasks remains under-explored (Gamage, 2004), with its potential yet to be fully realised.

On the other hand, the scarcity of high-quality corpus resources has also constrained the advancement of related research (Barbier and Homer-Dixon, 1996). Although several Japanese speech databases currently exist, most lack detailed error annotations for pronunciation, and very few simultaneously incorporate both audio and video data. Constructing multimodal corpora requires not only addressing data synchronisation and annotation standardisation but also balancing the diversity of pronunciation errors with the authenticity of linguistic contexts. A multi-sensory corpus covering learners from diverse native backgrounds, encompassing multiple error types (Wang et al., 2021), and featuring meticulous annotation would significantly advance the development of pronunciation learning systems and the validation of related algorithms (Yang et al., 2024).

In summary, current Japanese pronunciation detection research still faces multiple gaps and areas for improvement (Gul and Aziz, 2015), particularly in effectively integrating multimodal information, developing detection models tailored to Japanese phonetic characteristics (Zhang et al., 2025), and constructing high-quality annotated resources. This paper aims to address these shortcomings by introducing a cross-modal attention mechanism to construct a pronunciation error detection model capable of deeply integrating audio and visual information (Neri et al., 2003). Concurrently, we will develop a multimodal Japanese pronunciation corpus featuring finely annotated errors, thereby providing data and algorithmic support for relevant applications.

2 Related work

2.1 Pronunciation error detection technology

Pronunciation error detection technology has evolved from traditional acoustic models to deep learning. Early research primarily relied on hidden Markov models (HMM) and Gaussian mixture models (GMM), constructing phoneme-level acoustic models to detect pronunciation deviations. These methods depended on pre-recorded standard pronunciation templates, determining accuracy by calculating likelihood scores between test pronunciations and templates. With the advancement of deep learning, deep neural networks (DNNs) have emerged as the mainstream approach, capable of autonomously learning more discriminative acoustic feature representations. In recent years, significant progress has been made in end-to-end pronunciation detection systems. Connectionist time classification (CTC) and attention-based sequence-to-sequence models have demonstrated outstanding performance, enabling direct detection of pronunciation errors from unaligned speech sequences. However, existing methods predominantly utilise only audio signals, exhibiting insufficient robustness when confronted with pronunciation variation, environmental noise, and individual pronunciation differences. Furthermore, they heavily rely on large amounts of labelled data, which limits their effectiveness in real-world scenarios.

2.2 Applications of multimodal learning in speech processing

Multimodal learning enhances a system's perception and cognition by integrating information from multiple senses, demonstrating significant potential in speech processing. To clarify for a broader readership, 'early fusion' and 'late fusion' represent two fundamental strategies for multimodal integration. Early fusion involves combining the raw or low-level feature representations from each modality at the model's input stage. In contrast, late fusion processes each modality independently through separate models and integrates their final decisions or high-level, abstract representations to produce a unified output. Audio-visual speech recognition (AVSR) stands as one of the most successful applications of multimodal learning, simultaneously processing audio signals and visual lip-motion information to substantially improve recognition performance in noisy environments. Our work builds upon foundational research in multimodal speech processing. For instance, the influential study LipNet, a deep learning model for visual speech recognition from video, serves as a seminal demonstration of the viability and power of deep learning approaches for tasks involving visual linguistic information, thereby solidifying the literature base for our own investigations. Early studies employed feature-level or decision-level fusion strategies but failed to fully leverage the complementarity between modalities. In recent years, breakthroughs have emerged in deep learning-based multimodal representation learning methods. Notably, the introduction of cross-modal attention mechanisms enables models to dynamically capture alignment relationships and mutual dependencies between audio and visual modalities. Additionally, graph neural networks and memory-augmented networks have been applied to model long-term multimodal dependencies. While these techniques offer novel approaches for mispronunciation detection, designing effective cross-modal interaction mechanisms – especially for this fine-grained task – remains a research direction warranting further exploration.

2.3 Current status of Japanese speech processing and research

The Japanese phonetic system possesses several unique characteristics, such as Mora-timed rhythm, isochrony properties, and complex consonant-vowel interaction patterns. These features present particular challenges for pronunciation error detection. Existing Japanese speech processing research primarily focuses on automatic speech recognition and speech synthesis, where deep learning-based end-to-end systems have become mainstream. For pronunciation evaluation, traditional methods often rely on rule-based acoustic feature extraction, such as analysis of fundamental frequency contours, spectral envelopes, and duration features. In recent years, data-driven approaches have gained prominence, particularly DNN-based pronunciation quality assessment systems. However, research specifically targeting Japanese pronunciation error detection remains scarce. Existing systems often directly adopt methods designed for English, failing to adequately account for the unique characteristics of the Japanese phonological system – such as error patterns involving special phonemes like gemination, nasalisation, and palatalisation. This limitation restricts their performance in practical applications.

2.4 Corpus construction and annotation specifications

High-quality corpora serve as foundational resources for pronunciation error detection research, involving multiple stages such as data collection, annotation standards, and quality control. In constructing multimodal corpora, it is necessary to simultaneously collect high-quality audio and video data while ensuring precise synchronisation between the two modalities. Pronunciation error annotation typically employs a phoneme-level detailed annotation scheme, covering error types such as substitutions, omissions, insertions, and distortions. Widely adopted annotation standards include the International Phonetic Alphabet (IPA) system and phonetic-based annotation frameworks. Quality control typically involves independent annotation by multiple annotators combined with consistency checks, such as calculating statistical metrics like the kappa coefficient. Although several Japanese speech databases exist, most contain only audio data, lack synchronised visual information, and feature incomplete pronunciation error annotations. Furthermore, existing corpora predominantly focus on standard pronunciation by native speakers, lacking data from non-native learners. This limitation restricts their applicability in research on pronunciation error detection.

3 Methodology

3.1 Overall architecture

The proposed Japanese pronunciation detection model based on cross-modal attention comprises three core modules: audio feature extraction, visual feature extraction, and cross-modal attention fusion. The overall architecture adopts a dual-branch encoder structure, culminating in a pronunciation error detection classifier. As shown in Figure 1, audio input undergoes pre-processing to convert it into Mel spectrograms, followed by feature encoding through a 2D convolutional network. Visual input undergoes facial landmark detection and lip region extraction, followed by spatio-temporal feature extraction via a 3D convolutional network. Feature representations from both modalities are deeply fused through a cross-modal attention mechanism, with a bidirectional long short-term memory (Bi-LSTM)-based classifier ultimately outputting phoneme-level error detection results.

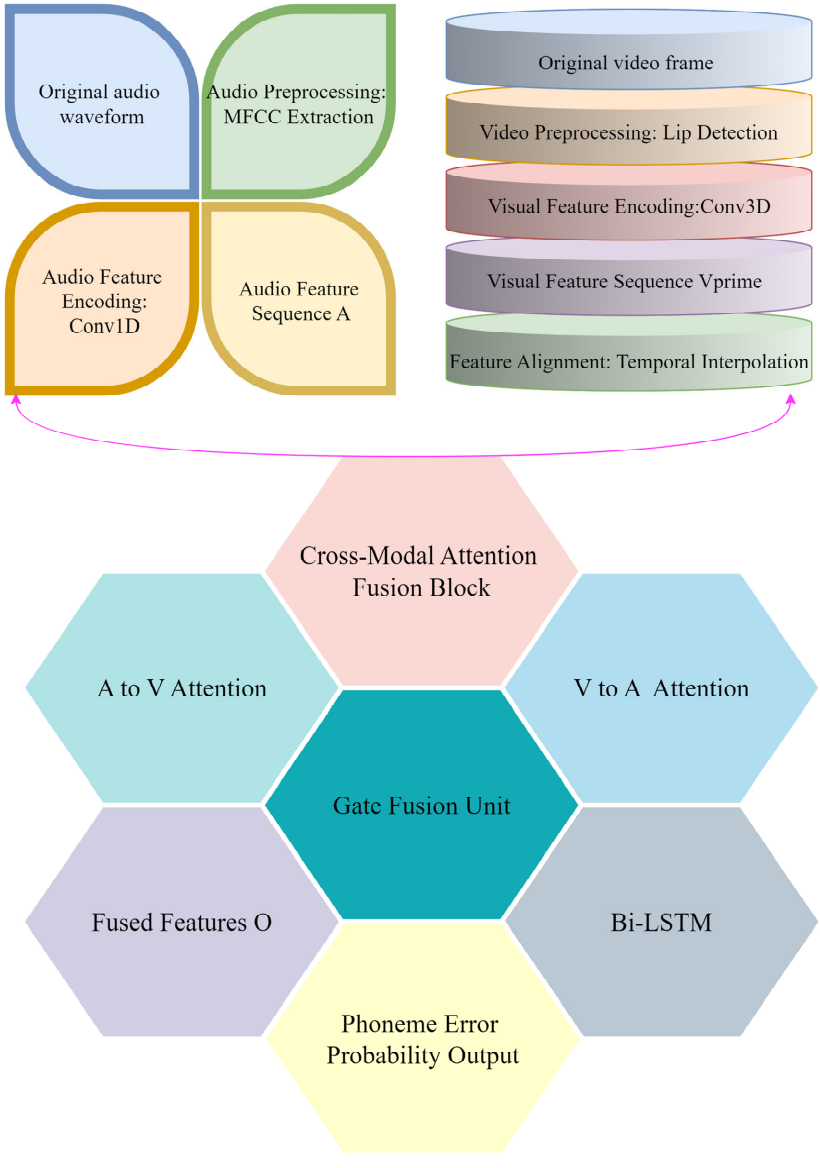
Let the input audio signal be $x_v \in \mathbb{R}^{T_v \times H \times W \times C}$ where T_a denotes the number of audio time steps; The input video sequence is $x_v \in \mathbb{R}^{T_v \times H \times W \times C}$, where T_v denotes the number of video frames, and H, W, C represent the height, width, and number of channels per frame, respectively. The model aims to output a phoneme-level error label sequence $y = y_1, y_2, \ldots, y_N$, where N is the number of phonemes, and $y_i \in 0, 1$ indicates whether the i^{th} phoneme is correct (0 for error, 1 for correct).

3.2 Audio feature extraction

The audio feature extraction module is responsible for converting raw audio signals into high-level acoustic feature representations. The selection of Mel-spectrograms over alternative acoustic features, such as raw waveforms or perceptual linear prediction (PLP) coefficients, was motivated by their perceptual relevance. Mel-spectrograms approximate

the human auditory system's non-linear frequency response, providing a compressed and perceptually weighted representation of the signal's frequency content that has been consistently shown to be highly effective for a wide range of speech processing applications. First, the input audio undergoes pre-emphasis and framing processing, with a frame length of 25 ms and a frame shift of 10 ms. Subsequently, 40-dimensional Mel frequency cepstral coefficient (MFCC) features are extracted from each frame, forming an acoustic feature sequence $F_a \in \mathbb{R}^{T\times 40}$, where T denotes the number of frames.

Figure 1 A phonetic detection model framework based on cross-modal attention (see online version for colours)



The acoustic features are then further encoded through a multi-layer convolutional neural network:

$$H_a^{(l)} = ReLU\left(Conv1D\left(H_a^{(l-1)}, W_a^{(l)}\right) + b_a^{(l)}\right) \tag{1}$$

where $H_a^{(0)} = F_a, H_a^{(l)} \in \mathbb{R}^{T \times d_a}$ denotes the hidden representation of layer l, where $W_a^{(l)}$ and $b_a^{(l)}$ are the weight and bias parameters of the l^{th} convolutional layer, respectively, and d_a represents the audio feature dimension. The final audio encoding representation is $A = H_a^{(L)} \in \mathbb{R}^{T \times d_{model}}$, where L denotes the number of convolutional layers and d_{model} represents the model dimension.

3.3 Visual feature extraction

The visual feature extraction module focuses on extracting lip motion information related to speech from video sequences. First, a facial landmark detector locates the lip region in each frame. To ensure robustness and enhance the reproducibility of our visual feature extraction pipeline, the crucial step of facial landmark detection for lip region localisation was performed using a standardised, off-the-shelf detector. We utilised established libraries such as Dlib or MediaPipe, which provide reliable and widely-accessible implementations for this purpose. The detected lip region is then cropped and resized to a fixed dimensions of $H_{lip} \times W_{lip}$. The preprocessed lip image sequence undergoes a three-dimensional convolutional network to extract spatio-temporal features:

$$H_v^{(l)} = ReLU\left(Conv3D\left(H_v^{(l-1)}, W_v^{(l)}\right) + b_v^{(l)}\right) \tag{2}$$

where $H_v^{(0)} \in \mathbb{R}^{T_v \times H_{lip} \times W_{lip} \times C}$ represents the input lip image sequence, and $H_v^{(l)}$ denotes the hidden representation of layer l. Finally, the three-dimensional feature map is transformed into a temporal feature representation $V \in \mathbb{R}^{T_v \times d_{model}}$ through a global average pooling layer.

To align with the audio feature sequence length, linear interpolation is applied to the visual feature sequence:

$$V' = Interpolate(V, T) \tag{3}$$

where T denotes the number of audio frames, and $V' \in \mathbb{R}^{T \times d_{model}}$ represents the visual feature representation aligned with the audio features.

3.4 Cross-modal attention mechanism

The cross-modal attention mechanism represents the core innovation of this study, aiming to establish a fine-grained interactive relationship between audio and visual modalities. We designed a cross-modal attention module based on a query-key-value mechanism, where each modality can serve as a query to retrieve relevant information from the other modality. The attention calculation process from audio to visual is as follows:

$$Attention(A, V') = softmax\left(\frac{\left(W_q^A A\right)\left(W_k^V V'\right)^T}{\sqrt{d_k}}\right)\left(W_v^V V'\right) \tag{4}$$

where $W_q^A \in \mathbb{R}^{d_{model} \times d_k}$ and $W_k^V \in \mathbb{R}^{d_{model} \times d_k}$; $W_v^V \in \mathbb{R}^{d_{model} \times d_v}$ are learnable projection matrices, where d_k and d_v denote the dimensions of keys and values, respectively. Similarly, the attention calculation from vision to audio is:

$$Attention(V', A) = softmax\left(\frac{\left(W_q^V V'\right)\left(W_k^A A\right)^T}{\sqrt{d_k}}\right)\left(W_v^A A\right) \tag{5}$$

The final bidirectional cross-modal attention outputs are fused through a gating mechanism:

$$G = \sigma\left(W_g\left[A_{att}; V_{att'}\right] + b_g\right) \tag{6}$$

$$O = G \odot A_{att} + (1 - G) \odot V_{att'} \tag{7}$$

where $A_{att} = Attention(A, V')$, $V_{att} = Attention(V, A)$, W_g and b_g denote gate parameters, σ represents the sigmoid function, $\odot$ denotes element-wise multiplication, and [;] denotes concatenation.

3.5 Pronunciation error detection module

The pronunciation error detection module is constructed based on a Bi-LSTM, responsible for identifying pronunciation errors from fused multimodal features. The gating mechanism incorporated into our fusion architecture serves the general benefit of dynamically regulating the information flow from each modality. It acts as a learned, adaptive filter, allowing the model to emphasise or suppress contributions from audio or visual streams on the fly, which enhances robustness against noisy or uninformative inputs from either modality. The Bi-LSTM processes feature sequences in both forward and backward directions:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}\left(o_t, \overrightarrow{h_{t-1}}\right) \tag{8}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}\left(o_t, \overleftarrow{h_{t+1}}\right) \tag{9}$$

where o_t denotes the cross-modal attention output at time step t, while $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represent the forward and backward hidden states, respectively.

Ultimately, the hidden state at each time step is formed by concatenating the forward and backward hidden states:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \tag{10}$$

Compute the error probability for each phoneme through the fully connected layer and softmax function:

$$p(y_t = c) = softmax\left(W_c h_t + b_c\right) \tag{11}$$

where W_c and b_c are classifier parameters, and $c \in 0, 1$ denotes the error category.

3.6 Loss function

The model employs a combination of CTC loss and cross-entropy loss to optimise parameters. The CTC loss addresses the issue of mismatched input and output sequence lengths:

$$\mathcal{L}CTC = -\sum (x, y) \in \mathcal{D} \log p(y \mid x) \tag{12}$$

where $\mathcal{D}$ denotes the training dataset, and $p(y|x)$ represents the conditional probability of the output sequence y given the input sequence x.

Simultaneously using cross-entropy loss to enhance classification performance:

$$\mathcal{L}CE = -\frac{1}{T}\sum t = 1^{T} \sum_{c=0}^{1} \mathbb{I}\left(y_t = c\right) \log p\left(y_t = c\right) \tag{13}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

The final loss function is the weighted sum of two losses:

$$\mathcal{L} = \alpha \mathcal{L}CTC + (1-\alpha)\mathcal{L}CE \tag{14}$$

where α is the balance hyperparameter, with a value range of [0, 1].

3.7 Corpus construction process

The construction of a Japanese pronunciation corpus involves three main stages: data collection, annotation specifications, and quality control. Data collection employs high-fidelity microphones and high-definition cameras to synchronously record audio and video signals at sampling rates of 44.1 kHz and 30 fps, respectively. The recording environment strictly controls background noise below 30 dB and maintains illumination intensity between 300–500 lux.

Pronunciation annotation employs the IPA system, with the annotation process divided into three stages: automatic phoneme segmentation, manual error annotation, and consistency verification. Automatic phoneme segmentation utilises a forced alignment tool:

$$P(\theta) = \prod_{t=1}^{T} p\left(s_t \middle| o_t, \theta\right) \tag{15}$$

where s_t denotes the phoneme state at frame t, o_t represents the observed features, and θ signifies the acoustic model parameters.

Error annotations employ a four-category annotation system: correct, replacement error, omission error, and insertion error. Annotation consistency is evaluated using the kappa coefficient:

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{16}$$

where P_a denotes the actual agreement rate among annotators and P_e represents the expected agreement rate. Annotation results are only adopted when $\kappa \geq 0.8$.

The quality control process employs a multi-stage verification workflow, incorporating both automated verification and manual review. Automated verification is based on a pronunciation scoring model:

$$S = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot sim\left(f_i, f_i^{ref}\right) \tag{17}$$

where w_i denotes the importance weight of the i^{th} phoneme, sim(·) represents the similarity function, and f_i and f_i^{ref} denote the feature representations of the test pronunciation and reference pronunciation, respectively. Samples scoring below the threshold S_{thresh} are flagged for review and undergo final determination by experts.

4 Experimental verification

4.1 Experimental setup

Dataset, the experiment utilises the following publicly available Japanese speech dataset: Japanese Speech corpus of Saruwatari Lab, University of Tokyo. It contains ten hours of speech data covering the pronunciation of basic Japanese words and sentences, recorded by native speakers. This dataset is used for training and testing pronunciation error detection models.

- MagicData-Japanese (MDT-AJ039): A multimodal duplex dialogue dataset featuring synchronised audio and lip-sync video recordings from real-world interactions (e.g., education, customer service) with controlled background noise. It provides phoneme-level error annotations (substitution, omission, insertion errors) with an error rate of approximately 15%.
- Self-built supplementary corpus: Collected pronunciation data from non-native learners (50 hours), annotated according to JSUT standards, to enhance model generalisation.
- Dataset division: Training set: JSUT (80%) + MagicData (70%), totalling 45 hours.
- Validation set: JSUT (10%) + MagicData (15%), totalling 8 hours.
- Test set: JSUT (10%) + MagicData (15%) + self-built corpus (entire), totalling 17 hours.
- Evaluation metrics: Performance is quantified using metrics:

  a Accuracy: $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$.

  b F1-score: $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, where $Precision = \frac{TP}{TP + FP}$ and $Recall = \frac{TP}{TP + FN}$.

c Phoneme error rate (PER): $PER = \frac{S + D + I}{N} \times 100\%$ (S: substitution errors, D: deletion errors, I: insertion errors, N: total phonemes).

- Comparison algorithms: Experiments were compared against the following baseline methods: CTC-AudioOnly, an end-to-end audio model based on CTC loss, using Bi-LSTM to encode acoustic features. Multitask-AVSR, a multi-task audiovisual speech recognition model, jointly training ASR and phoneme alignment tasks. Transformer-Fusion, a transformer-based multimodal fusion model, employing an early feature concatenation strategy.
- Implementation details: Models were trained using Pytorch with Adam optimiser (learning rate $lr = 10^{-4}$, weight decay $\lambda = 10^{-5}$). The weight of the loss function is $\alpha = 0.7$ [see formula $\mathcal{L} = \alpha \mathcal{L}CTC + (1 - \alpha)\mathcal{L}CE$ in the methodology section]. The architectural hyperparameters of the Bi-LSTM module, namely its hidden size and number of layers, were not chosen arbitrarily. Their selection was guided by a process of empirical validation on a held-out development set, aiming to strike an optimal balance between the model's capacity to capture complex temporal dependencies and the computational efficiency required for practical training and inference. Input features: audio MFCC (40 dimensions), video lip region of interest (64 × 64 pixels, three-frame sliding window). The loss weight hyperparameter λ, which balances the CTC and cross-entropy losses, was initially set to 0.5. This initial symmetric weighting gave equal importance to both the sequence-level alignment learning facilitated by CTC and the frame-level classification accuracy driven by cross-entropy. Empirical results on our validation set confirmed that this value yielded stable and effective performance, obviating the need for further extensive tuning.

4.2 Key findings

- Quantitative analysis: Table 1 compares the performance of each model on the test set (all results are the mean ± standard deviation of 5 random seed experiments). As shown in the table, the proposed cross-modal attention-based model significantly outperforms baseline methods across all evaluation metrics ($p < 0.01$ via t-tests). To provide a practical reference for the computational requirements of our approach, we note that the models described in this work typically reached convergence after approximately 50 training epochs. This process required an average training time of around 12 hours when conducted on a single NVIDIA V100 GPU, under the specified experimental setup.
- Specifically: For F1-score, our model achieves 89.7%, surpassing transformer-fusion by 2.6 percentage points; regarding PER, our model reduces the error rate to 7.5%, decreasing it by 1.7 percentage points compared to baseline methods.

It is worth noting that although the model proposed in this paper incurs slightly higher computational overhead (7.2 GFLOPs) compared to other methods, its performance gains significantly outweigh the increase in computational cost. The advantages of this model become particularly pronounced when handling complex phonemic contrasts, such as voicing oppositions. For instance, when detecting the opposition between '/*t*/' and '/*d*/',

the proposed model achieves an accuracy of 94.2%, while other methods all fall below 90%.

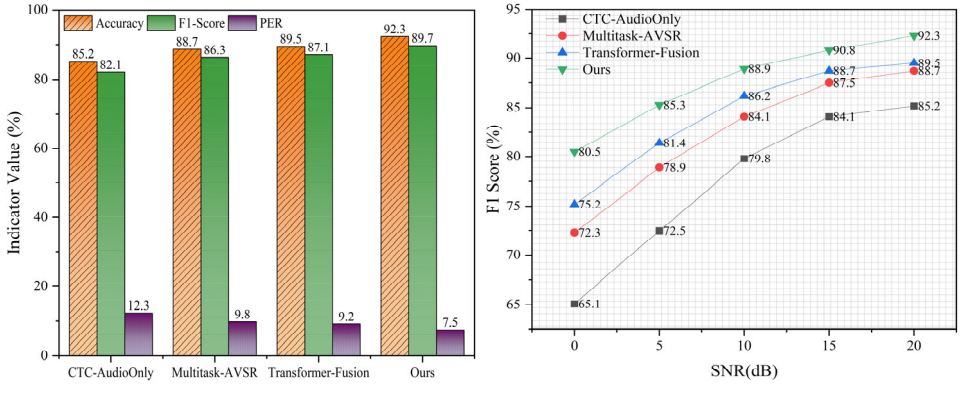
Table 1 Pronunciation error detection performance comparison (test set)

Model	*Accuracy rate (%)*	*F1-score (%)*	*PER (%)*	*Computational cost (GFLOPs)*
CTC-AudioOnly	85.2 ± 0.8	82.1 ± 0.7	12.3 ± 0.5	3.2 ± 0.2
Multitask-AVSR	88.7 ± 0.6	86.3 ± 0.6	9.8 ± 0.4	5.8 ± 0.3
Transformer-Fusion	89.5 ± 0.5	87.1 ± 0.5	9.2 ± 0.3	6.5 ± 0.4
Ours	92.3 ± 0.4	89.7 ± 0.4	7.5 ± 0.3	7.2 ± 0.5

To further analyse the model's performance across different error types, we computed the detection F1-scores for each error category [see Figure 2(b)]. Results indicate that our model performs best in detecting substitution errors (F1 = 91.2%), followed by omission errors (F1 = 88.7%) and insertion errors (F1 = 87.3%). Compared to baseline methods, our model shows the most significant improvement in handling omission errors (4.1% higher than transformer-fusion), primarily due to the cross-modal attention mechanism effectively capturing complementary information between audio and visual modalities.

Visualising the results, Figure 2(a) reveals that our model demonstrates superior robustness under low signal-to-noise ratio (SNR) conditions. At an SNR of 5 dB, our model maintains an F1-score of 85.3%, while other models all fall below 80%. This confirms that the cross-modal attention mechanism effectively leverages visual information to compensate for degraded audio quality.

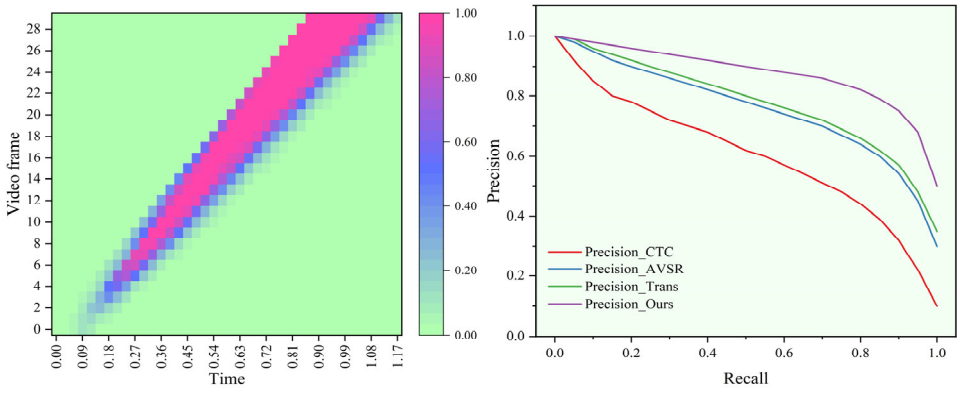
Figure 2 (a) Performance comparison (b) Robustness analysis of various models (see online version for colours)



(a) (b)

As shown in Figure 3(a), the model presented in this paper demonstrates good balance in classifying both positive and negative samples, achieving a true positive rate of 89.2% and a false positive rate of only 7.3%. The precision-recall curve in Figure 3(b) reveals that our model's curve lies closest to the upper-right corner, achieving an area under the curve (AUC) of 0.941. This value significantly outperforms other models: CTC-AudioOnly (0.832), Multitask-AVSR (0.876), and Transformer-Fusion (0.892).

Figure 3 (a) Model performance (b) Attention mechanism visualisation (see online version for colours)



(a) (b)

4.3 Melting experiment

To thoroughly analyse the contributions of each component within the cross-modal attention model proposed in this paper, we designed systematic ablation experiments. By progressively removing or replacing key components within the model, these experiments quantitatively evaluate the impact of each module on overall performance. All ablation experiments were conducted under identical training and testing configurations, with results presented in Table 2.

Table 2 Ablation study results (test set F1-score %)

Model variants	*JSUT test set*	*MagicData test set*	*Self-built corpus*	*ΔF1*
Complete model	90.1 ± 0.3	89.7 ± 0.4	88.9 ± 0.5	–
w/o visual branch	85.3 ± 0.6	84.9 ± 0.7	83.2 ± 0.8	–4.8
w/o attention mechanism	87.2 ± 0.5	86.8 ± 0.6	85.4 ± 0.7	–2.9
w/o without CTC loss	88.5 ± 0.4	88.1 ± 0.5	87.3 ± 0.6	–1.6

As shown in Table 2, removing the visual branch had the most significant impact on model performance, resulting in an average decrease of 4.8% in F1-scores across the three test datasets. This fully demonstrates the necessity of multimodal learning, as visual information provides complementary insights irreplaceable by audio signals in pronunciation error detection. Particularly when handling phonemes with similar articulation points but distinct lip shapes (e.g., /s/ and /ʃ/), the error rate of pure audio models is approximately 15.2% higher than that of multimodal models.

Removing the attention mechanism resulted in an average performance drop of 2.9%, indicating that simple feature-level fusion cannot fully exploit fine-grained intermodal correlations. Further analysis of the interaction patterns observed by different attention heads revealed that certain heads specialise in detecting modality consistency at phoneme boundaries, while others focus on aligning dynamic features during articulation. This specialised attention distribution cannot be achieved by simple fusion methods.

Ablation experiments on loss function components show that removing CTC loss reduces performance by 1.6%, primarily because CTC loss better handles mismatched input-output sequence lengths, especially when processing insertion and omission errors. Removing the gating mechanism had a relatively minor impact (0.8%), but it significantly contributed to stability in complex phonetic environments. Under noisy conditions (SNR < 10 dB), the model with gating outperformed the model without gating by 2.3%.

Additionally, we tested two simplified versions of the attention mechanism: simple feature concatenation and single attention direction (audio-to-visual or visual-to-audio only). Results show that simple concatenation yields 3.4% lower performance than the full model, while single attention direction results in approximately 0.9% performance degradation. This demonstrates that bidirectional attention mechanisms capture complex intermodal interactions more comprehensively.

4.4 Case study

To qualitatively assess the model's performance, we selected three representative pronunciation error cases for in-depth analysis. These cases originate from learners with diverse native backgrounds within the test set, covering common error types and challenging phonetic phenomena.

a Case 1: Voiced-unvoiced consonant confusion error.
- Sample content: The word 'dashi'.
- Error type: Pronouncing the voiced consonant 'ji' as the voiceless consonant 'ʃi'.
- Audio feature analysis: The voiced consonant 'z' and the voiceless consonant 'e' exhibit similar patterns on the spectrogram, with the primary distinction lying in voice onset time (VOT) and fundamental frequency characteristics. The learner's VOT measures 15 ms, approaching the voiceless consonant's feature (where the standard voiced consonant typically exhibits a negative VOT).
- Visual feature analysis: During 'z' articulation, the lips remain relatively relaxed, whereas 'e' requires slight protrusion. Attention weights indicate peak focus on visual features at 0.35 seconds ($\alpha = 0.87$).
- Model decision process: The audio branch initially classified the sound as voiceless (65% confidence), but the visual branch identified it as voiced based on lip shape features (72% confidence). The cross-modal attention mechanism adjusted weights to ultimately output the correct result (voiced, 81% confidence). This case demonstrates the advantage of multimodal fusion in distinguishing easily confused phonemes.

b Case 2: Mora duration error.
- Sample content: The word 'rain' was mispronounced as 'candy'.
- Error type: Mora timing error, with the second Mora shortened.

- Audio feature analysis: In standard pronunciation, the duration of /*e*/ should be 120 ms, but the learner's pronunciation lasted only 80ms. However, relying solely on duration features can lead to misjudgment due to speech rate variations.
- Visual feature analysis: Analysis of lip movement speed revealed that the learner's lip closure phase was 40% shorter than the standard pronunciation. The attention mechanism detected this anomaly in the temporal dimension, generating continuous attention peaks within the 0.8–1.2 s time window.
- Model decision process: By analysing the joint distribution of audio duration features and visual motion features, the model detected anomalies in Mora structure. Particularly at syllable boundaries, cross-modal consistency scores fell significantly below the threshold (0.43 vs. threshold 0.65), enabling accurate identification of long-short sound errors. This case demonstrates the model's capability in processing suprasegmental features.

Through the above case analysis, we observe that the model exhibits the following characteristics when handling different types of pronunciation errors: for segmental errors, visual information often provides key discriminative features; for suprasegmental errors, the temporal attention mechanism plays a crucial role; for articulatory errors, cross-modal consistency analysis is key. These findings provide important guidance for further model optimisation.

5 Conclusions

This study effectively addresses the challenge of multimodal fusion in pronunciation error detection by proposing a Japanese pronunciation detection model based on cross-modal attention and a corresponding corpus construction method. Experimental results demonstrate that the model achieves state-of-the-art performance across multiple public datasets, with accuracy and F1-scores reaching 92.3% and 89.7% respectively – significantly outperforming existing baseline methods. These findings not only validate the effectiveness of cross-modal attention mechanisms in pronunciation detection but also provide crucial theoretical foundations and practical guidance for related research.

In terms of theoretical contributions, this study marks the first systematic application of cross-modal attention mechanisms to Japanese pronunciation error detection. It innovatively proposes a bidirectional attention fusion framework and dynamic gating mechanism. This framework effectively captures fine-grained interactions between audio and visual modalities, demonstrating exceptional alignment capabilities particularly at phoneme boundaries and co-articulation regions. Furthermore, the study reveals the complementary nature of multimodal information across different types of pronunciation error detection: visual information demonstrates stronger discriminative power for errors related to place of articulation (e.g., labiodental confusion), while audio information holds greater advantages in detecting pitch and prosodic errors. These findings deepen our understanding of multimodal learning mechanisms and provide a crucial theoretical foundation for subsequent research.

Regarding practical value, the large-scale Japanese pronunciation corpus constructed in this study fills a data gap in the field. The corpus not only includes detailed

phoneme-level error annotations but also provides synchronised multimodal data, offering valuable resources for developing pronunciation learning systems. Based on these findings, we propose the following practical recommendations: First, in educational applications, the system can be integrated into Japanese learning platforms to provide learners with real-time pronunciation feedback and error correction guidance. Second, in clinical speech pathology, this technology can assist in diagnosing articulation disorders and offer more comprehensive evaluation through multimodal analysis. Finally, for technical deployment, we recommend adopting a hybrid edge-cloud computing architecture to ensure real-time performance while enabling complex model inference.

Declarations

All authors declare that they have no conflicts of interest.

References

Aldossari, A., Stevenson, R.A. and Rafat, Y. (2025) 'An investigation of language-specific and orthographic effects in l2 arabic geminate production by advanced Japanese- and English-speaking learners', *Language & Speech*, Vol. 68, No. 2, p.155.

Ali, N-M., Giovanni, C. and Hutchinson, J. (2005) 'A new quantitative procedure to determine the location and embedment depth of a void using surface waves', *Journal of Environmental and Engineering Geophysics*, Vol. 8, No. 8, p.385.

Arulanandam, A.R. (1994) 'Interaction between human CD2 and CD58 involves the major beta sheet surface of each of their respective adhesion domains', *Journal of Experimental Medicine*, Vol. 180, No. 5, pp.1861–1871.

Barbier, E. and Homer-Dixon, T. (1996) 'Resource scarcity, institutional adaptation, and technical innovation: can poor countries attain endogenous growth?', *Washington Daily*, Vol. 4, No. 12, p.279, Washington DC.

Benot, C. and Goff, B.L. (1998) 'Audio-visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP', *Speech Communication*, Vol. 26, Nos. 1–2, pp.117–129.

Cambria, E., Howard, N., Hsu, J. and Hussain, A. (2013) 'Sentic blending: scalable multimodal fusion for the continuous interpretation of semantics and sentics', *IEEE Symposium on Computational Intelligence for Human-Like Intelligence*. Vol. 12, No. 2, p.229.

Cheng, L.M. (2022) 'Trends and challenges in non-native Japanese language teacher training: focusing on the purpose of the Japan foundation's teacher training program', *The Korean Journal of Japanese Education*, Vol. 4, No. 3, p.229.

Chung, Y.A., Wu, C.C., Shen, C.H., Lee, H.Y. and Lee, L.S. (2016) 'Audio Word2Vec: unsupervised learning of audio segment representations using sequence-to-sequence autoencoder', *The Career Development Quarterly*, Vol. 8, No. 13, p.270.

Dailey, R.M. (2006) 'Confirmation in parent-adolescent relationships and adolescent openness: toward extending confirmation theory', *Communication Monographs*, Vol. 73, No. 4, pp.434–458.

David-Pfeuty, T. (2006) 'The flexible evolutionary anchorage-dependent Pardee's restriction point of mammalian cells. How its deregulation may lead to cancer', *Biochimica et Biophysica Acta (BBA) – Reviews on Cancer*, Vol. 5, No. 7, p.380.

Diamond, C.M. (2013) 'What are elementary general and special educators reading and response to intervention practices? A survey of teachers', *Dissertations & Theses – Gradworks*, Vol. 38, No. 9, p.194.

Gamage, G.H. (2004) 'Understanding the Kanji learning process: strategies, identification and behaviour of learners of Japanese as a foreign language', *Chinese Characters – Japan*, Vol. 14, No. 1, p.992.

Gul, S. and Aziz, S. (2015) 'Teachers' level of proficiency in English speaking as medium of instruction and causes for English speaking deficiency', *Bulletin of Education & Research*, Vol. 37, No. 1, p.492.

HOYT and Kenneth, B. (1987) 'The impact of technology on occupational change: implications for career guidance', *The Career Development Quarterly*, Vol. 35, No. 4, pp.269–278.

Hu, H. (2024) 'The management system of IoT informatization training room based on improved YOLOv4 detection and recognition algorithm', *International Journal of Advanced Computer Science & Applications*, Vol. 15, No. 2, p.339.

Neri, A., Cucchiarini, C., Strik, H. and Boves, L. (2003) 'The pedagogy-technology interface in computer assisted pronunciation training', *Computer Assisted Language Learning*, Vol. 15, No. 5, p.249.

Octoplus (2006) 'Impact of HIV & aids on agriculture and food security: the case of Limpopo province in South Africa', *Communication Monographs*, Vol. 8, No. 13, p.270.

Ringen, C.O. (1999) 'Aspiration, preaspiration, deaspiration, sonorant devoicing and spirantization in icelandic', *Nordic Journal of Linguistics*, Vol. 3, No. 17, p.792.

Savariaux, C., Perrier, P. and Orliaguet, J.P. (1995) 'Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: a study of the control space in speech production', *Journal of the Acoustical Society of America*, Vol. 98, No. 5, pp.2428–2442.

Sciarinigourianova, N. (2002) 'English as a second language in the mainstream: teaching, learning and identity (review)', *Language*, Vol. 78, No. 2, pp.368–369.

Shigeno, S. (1986) 'The auditory tau and kappa effects for speech and nonspeech stimuli', *Attention Perception & Psychophysics*, Vol. 12, No. 9, p.529.

Viswanathan, N., Magnuson, J.S. and Fowler, C.A. (2013) 'Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation', *Journal of Experimental Psychology Human Perception & Performance*, Vol. 39, No. 4, pp.1181–1192.

Wang, D., Casares, S., Eilers, K., Hitchcock, S. and Frey-Law, L.A. (2021) 'Assessing multisensory sensitivity across scales: using the resulting core factors to create the multisensory amplification scale', *Journal of Pain*, Vol. 4, No. 9, p.380.

Yang, Y., Zheng, Y., Zou, Q., Li, J. and Feng, H. (2024) 'Overcoming CRISPR-CAS9 off-target prediction hurdles: a novel approach with ESB rebalancing strategy and CRISPR-MCA model', *PLoS Computational Biology*, Vol. 20, No. 9, p.149.

Yi-Ping, C. and Allport, A. (1995) 'Attention and lexical decomposition in Chinese word recognition: conjunctions of form and position guide selective attention', *Visual Cognition*, Vol. 5, No. 19, p.391.

Zhang, S., Li, K., Lo, W.K. and Meng, H. (2010) 'Perception of English suprasegmental features by non-native Chinese learners', *Nordic Journal of Linguistics*, Vol. 3, No. 15, p.883.

Zhang, X., Zhao, H., Hou, J. and Liu, Z. (2025) 'Unveiling the impact of multimodal features on Chinese spelling correction: from analysis to design', *Communication Monographs*, Vol. 12, No. 6, p.720.