# Comparative Analysis of Japanese Speech: Applying Dynamic Time Warping and Precise Word Segmentation for Pronunciation Assessment

Supaporn Bundasak[1,*], Kollathee Wisawayotanan[1], Chen Chien-Chang[2]

[1]Department of Computer Science and Information Technology, Faculty of Science, Sriracha, Kasetsart University, Sriracha Campus, Thailand

[2]Department of Computer Science and Information Engineering, Tamkang University, New Taipei City, Taiwan

Received 04 January 2026; received in revised form 22 April 2026; accepted 24 April 2026

DOI: https://doi.org/10.46604/ijeti.2026.16049

## Abstract

To address the limitations of conventional pronunciation assessment in handling acoustic variability. This research presents a novel framework for Japanese pronunciation assessment that integrates self-supervised speech representations with temporal alignment to facilitate granular feedback. The proposed methodology utilizes Wav2Vec 2.0 for automated, high-precision word segmentation, followed by dynamic time warping (DTW) to quantify similarity in pitch-accent patterns. Experimental results indicate that the long short-term memory (LSTM)-based classification model achieves an accuracy of 92.5% with an F1-score of 0.92, demonstrating high reliability in pronunciation discrimination. Furthermore, the system effectively isolates prosodic deviations through word-level distance heatmaps, providing actionable diagnostic feedback for learners. This study contributes a robust, model-driven pipeline that enhances the diagnostic capability of computer-assisted pronunciation training (CAPT) systems for Japanese language learning.

**Keywords:** Japanese characters, speech analysis, pronunciation assessment, dynamic time warping, word segmentation

## 1. Introduction

Japanese, spoken by approximately 125 million people, features a complex vocabulary and four-script writing system—Kanji, Hiragana, Katakana, and Romaji—shaped by centuries of Sino-Japanese borrowing and modern European influences [1]. For many learners, mastering Japanese pronunciation remains a formidable challenge due to its unique phonetic structure and distinct pitch-accent patterns. The absence of reliable self-evaluation mechanisms often results in persistent segmental and prosodic errors, which significantly impede intelligibility and long-term proficiency. To address these acquisition barriers, recent advancements in speech processing and machine learning have catalyzed the development of robust computer-assisted pronunciation training (CAPT) tools. These tools offer scalable solutions for automated diagnostic feedback.

A fundamental component of effective CAPT systems involves robust feature extraction and precise alignment techniques. For feature extraction, mel-frequency cepstral coefficients (MFCCs) are a standard due to their ability to approximate human auditory perception. Kiran (2021) presented a comprehensive framework for MFCC extraction, demonstrating that these features effectively capture both the spectral and temporal characteristics of phonemes, which is critical for accurate speech recognition [2]. This is supported by Abdul and Al-Talabani (2022), who reviewed MFCC applications and highlighted their robustness in speaker identification and emotion analysis tasks [3]. Furthermore,

* Corresponding author. E-mail address: supaporn.band@ku.th

comparative research by Kaur et al. [4] confirmed that MFCCs achieve the highest accuracy for speaker-dependent recognition. This performance is superior to techniques such as linear predictive coding (LPC) and perceptual linear prediction (PLP), particularly under varying noise conditions.

Dynamic time warping (DTW) is extensively utilized for aligning temporal sequences of varying lengths. It offers robust speaker discrimination, where temporal alignment often outweighs amplitude normalization in benchmarking learner utterances [5]. Studies indicate that integrating DTW with MFCCs can outperform contemporary deep learning models in recognition accuracy, despite remaining computational challenges [6]. Furthermore, hybrid frameworks fusing DTW-aligned pitch contours with goodness of pronunciation (GOP) scores surpass single-metric systems by providing concurrent feedback on phonemic and prosodic accuracy [7]. The efficacy of these methodologies is further supported by high-quality corpora such as the Japanese versatile speech (JVS) dataset, which provides essential resources for multi-speaker modeling and phonetic research [8].

Despite these methodological advancements, existing CAPT systems often focus on evaluating overall pronunciation scores rather than providing in-depth feedback on individual words. While techniques such as DTW and MFCCs are effective for global alignment, they are often lack insufficient granularity for learners to identify and correct specific segmental errors. Moreover, conventional utterance-level recognition usually struggles to pinpoint exact error locations due to inconsistent phonetic boundaries. This limitation leaves learners aware of their errors but unsure of how to rectify them.

Therefore, this research aims to develop a Japanese phonetic comparison process that enables instructors and learners to assess and improve pronunciation with greater precision. The proposed system utilizes a model-driven approach, incorporating Wav2Vec 2.0 for accurate automatic word segmentation. This is followed by the extraction of key acoustic features—including MFCCs, waveform, spectrogram, and pitch—which are processed using a DTW algorithm to calculate similarity scores. The primary novelty of this study lies in the synergistic integration of Wav2Vec 2.0-driven precise word-level segmentation with DTW to provide granular pronunciation feedback. Unlike existing systems that offer global utterance scores, the proposed framework isolates segmental errors through automated boundary detection, allowing for the generation of word-specific distance heatmaps. This approach enables a more targeted diagnostic tool for Japanese learners by specifically visualizing deviations in pitch-accent patterns at the lexical level. Finally, this process is designed to be deployed as a web application.

The remainder of this paper is organized as follows: Section 2 describes the theoretical background of sound analysis, Wav2Vec, and DTW. Section 3 details the methodology, including data collection, Wav2Vec, and model training. Section 4 presents the experimental results and discussion regarding speaker comparison and system performance. Finally, Section 5 concludes the study with directions for future research.

## 2. Theoretical Background

This section elucidates the fundamental theoretical principles and computational frameworks underpinning the proposed system. The discussion encompasses a wide spectrum of methodologies, ranging from initial acoustic feature extraction to the application of sophisticated deep learning models for high-precision word segmentation and temporal alignment. By synthesizing these theoretical components, this section establishes a robust technical foundation for the subsequent development and evaluation of the Japanese pronunciation assessment pipeline.

### *2.1. Acoustic feature analysis*

Speech signals are characterized as non-stationary processes, utilizing the short-time fourier transform (STFT) to identify resonances and transient events through high-resolution spectrograms [9-10]. To evaluate Japanese pronunciation, intensity and fundamental frequency (F0) are extracted, with F0 serving as the deterministic factor for pitch-accent patterns where tone

transitions distinguish lexical meaning [11]. Furthermore, MFCCs are integrated with these prosodic descriptors to establish a robust framework for phonetic discrimination. Finally, DTW is applied to effectively align and quantify similarity between native and learner utterances, providing standardized diagnostic feedback for pronunciation assessment.

### *2.2. Wav2vec 2.0 framework*

Wav2Vec is a pioneering self-supervised framework for deriving high-quality speech representations directly from raw audio waveforms, thereby substantially reduces dependence on labeled data. In the original formulation, raw speech is processed using a multi-layer convolutional encoder to yield low-level feature embeddings. These embeddings are subsequently quantized into discrete codebook entries. Additionally, a contrastive loss encourages embeddings drawn from the same temporal context to be more similar than those from different contexts. This mechanism enables the model to capture salient acoustic patterns without supervision [12]. Building on this foundation, Wav2Vec 2.0, which replaces the convolutional context network with a Transformer architecture. This version adopts a mask-and-predict objective analogous to bidirectional encoder representations from transformers (BERT), where randomly masked spans of latent speech representations must be predicted from unmasked context vectors. This innovation yields contextualized embeddings that achieve state-of-the-art word error rates even under noisy conditions when fine-tuned on modest amounts of labeled speech [13].

### *2.3. Dynamic time warping*

DTW is a dynamic programming algorithm utilized to quantify similarity between two time-series sequences of varying lengths through nonlinear temporal axis alignment. By constructing a cost matrix, DTW identifies an optimal warping path that minimizes cumulative alignment cost. This process is subject to boundary, continuity, and monotonicity constraints, which facilitate precise correspondence between sequences [14]. As illustrated in Fig. 1, this mechanism establishes a frame-by-frame mapping that effectively compensates for temporal distortions, unlike traditional linear alignment. In the context of pronunciation assessment, DTW is essential for mitigating natural variations in speaking rates and individual phonetic durations. This ensures that corresponding acoustic features are accurately aligned for robust similarity measurement [15-16].

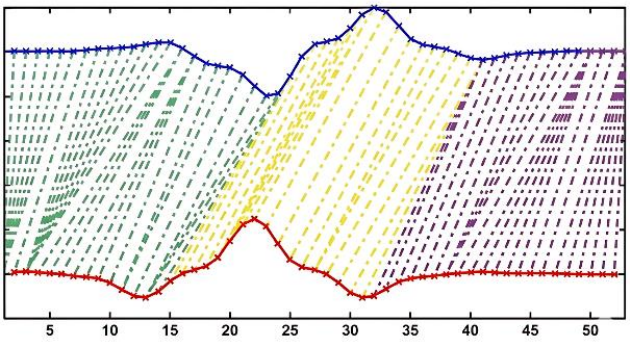


Fig. 1 Mechanism of DTW temporal alignment

Fig. 1 presents the representation of DTW alignment between two acoustic sequences. The dashed lines illustrate the optimal non-linear warping path $W = \{w_1, w_2, \ldots, w_K\}$, which establishes a frame-by-frame correspondence between the reference and test signals. This mechanism minimizes the cumulative distance D (n, m) by compensating for temporal distortions and variations in phonetic duration.

DTW formulation [14]. Let $X = (x_1,x_2,\ldots,x_n)$ and $Y = (y_1,y_2,\ldots,y_m)$ denote two feature sequences of length n and m, respectively, where each $x_i$ and $y_j$ is a feature vector (e.g., MFCCs). The local distance, explicitly denoted as d(i,j), represents the spatial dissimilarity between two feature vectors $x_i$ and $y_j$. It is formally defined using the euclidean metric as:

$$d(i, j) = \left| x_i - y_j \right| \tag{1}$$

where | · | denotes a suitable distance metric, typically the Euclidean distance. Subsequently, the elements D(i,j) of the accumulated cost matrix are defined to represent the minimal cumulative distance from the initial state (1,1) to the grid coordinate (i,j). This value is computed via the following recursive function:

$$D(i,j) = d(i,j) + min\begin{cases} D(i-1,j), (insertion) \\ D(i,j-1), (deletion) \\ D(i-1,j-1), (match) \end{cases} \quad (2)$$

To execute the dynamic programming recurrence, the elements of the accumulated cost matrix must be initialized according to the following boundary conditions. These equations define the cumulative cost at the starting point and along the edges of the alignment grid, ensuring that the path originates correctly at the first frame of both sequences:

$$D(1,1) = d(1,1) \quad (3)$$

$$D(i,1) = D(i-1,1) + d(i,1), \;\; i > 1 \quad (4)$$

$$D(1,j) = D(1,j-1) + d(1,j), \;\; j > 1 \quad (5)$$

The optimal DTW distance between sequences X and Y is then given by:

$$\mathrm{DTW(X,Y)} = \mathrm{D(n,m)} \quad (6)$$

where D(n,m) corresponds to the minimal cumulative cost along the optimal warping path from (1,1) to (n,m). This path satisfies both monotonicity and continuity constraints, ensuring a valid temporal alignment between the two sequences. In this framework, the feature vectors $x_i$ and $y_j$ represent multi-dimensional prosodic descriptors comprising MFCCs, fundamental frequency ($F_0$), and signal intensity. The normalized DTW distance is finally computed as D(n,m) / K, providing a scale-invariant metric to evaluate the pronunciation similarity between the learner and the native reference.

To formally define the alignment, let the optimal warping path be denoted as a sequence $W = \{w_1, w_2, \ldots, w_K\}$, where the k-th element $w_k = (i_k, j_k)$ maps the temporal index i of sequence X to index j of sequence Y. The path length K is bounded by $\max(n, m) \leq K < n + m - 1$. This warping path is strictly governed by three fundamental algorithmic constraints:

(1) Boundary condition: $w_1 = (1, 1)$ and $w_K = (n, m)$, ensuring the entire sequences are completely aligned.
(2) Monotonicity condition: $i_{k-1} <= i_k$ and $j_{k-1} <= j_k$, preserving the chronological progression of the acoustic features.
(3) Step-size condition: $w_k - w_{k-1} \in \{(1,0), (0,1), (1,1)\}$, ensuring continuity without skipping critical time frames.

Furthermore, computing the unconstrained DTW matrix yields a time and space complexity of $O(n \times m)$, which is computationally expensive and susceptible to pathological alignments (e.g., mapping a short consonant to an excessively long vowel). To mitigate this, a global warping constraint, specifically the Sakoe-Chiba band, is imposed on the search space. The alignment is restricted such that.

$$\left| i_k - j_k \right| \leq r \quad (7)$$

where *r* denotes the window size. This optimization ensures that the warping path stays within a fixed distance from the diagonal, effectively reducing the computational cost to $O(r \bullet \max(n,m))$ while preserving physically plausible phonetic mapping.

### *2.4. Specialization automatic pronunciation assessment (APA)*

APA systems utilize structured pipelines to benchmark learner speech against native references. Initial stages involve signal pre-processing for noise reduction and normalization to maintain signal integrity [17-18]. Subsequently, feature extraction transforms waveforms into spectral descriptors such as MFCCs [18-19], followed by temporal alignment using DTW to quantify rhythmic and prosodic deviations [18-19]. Traditional scoring mechanisms often employ the GOP metric within hidden markov model–deep neural network (HMM-DNN) architectures to aggregate frame-level posterior probabilities into comprehensive quality scores. Alternatively, contemporary end-to-end models, including connectionist temporal classification (CTC) and attention-based decoders, streamline this process by mapping raw audio directly to phonemic transcripts. These models frequently enhance performance in CAPT applications by eliminating modular dependencies [18, 20].

## 3. Methodology

Before detailing the methodological framework, the theoretical underpinnings and practical imperatives of word-level segmentation in Japanese speech processing are established. Often, conventional utterance-level recognition struggles with inconsistent phonetic boundaries and high computational demands. Accordingly, a model-driven approach is applied to partition continuous speech into discrete word units. This segmentation enables targeted extraction of acoustic features—specifically MFCCs, pitch, and signal intensity—for each segment and supports precise similarity assessments via DTW. Anchored by the objective of enhanced recognition accuracy, the resulting pipeline is designed to rigorously evaluate every component. This process extends from data acquisition and acoustic-model-based segmentation through DTW distance computation. Subsequent validation confirmed against manually annotated transcripts to ensure a robust and scalable solution for word-level Japanese speech recognition.

The methodology comprises four sequential stages. First, Japanese speech recordings from male and female native speakers are collected and manually transcribed to produce labeled audio files. Second, these recordings are processed using the Wav2Vec 2.0 framework to perform automatic transcription and word-level segmentation. Subsequently, key acoustic features—including MFCCs, pitch, and intensity—are extracted using the Librosa library. Third, DTW (via the fastdtw or dtaidistance implementation) is applied to compute normalized similarity scores between each segmented word in the training set and its corresponding reference segment. Finally, the long short-term memory (LSTM)-based classification model is trained on DTW-derived feature representations. This model is evaluated using word-level accuracy, F1 Score, and mean squared error, with performance trends visualized as heatmaps. Fig. 2 shows the flowchart for the research approach, which consists of six steps.

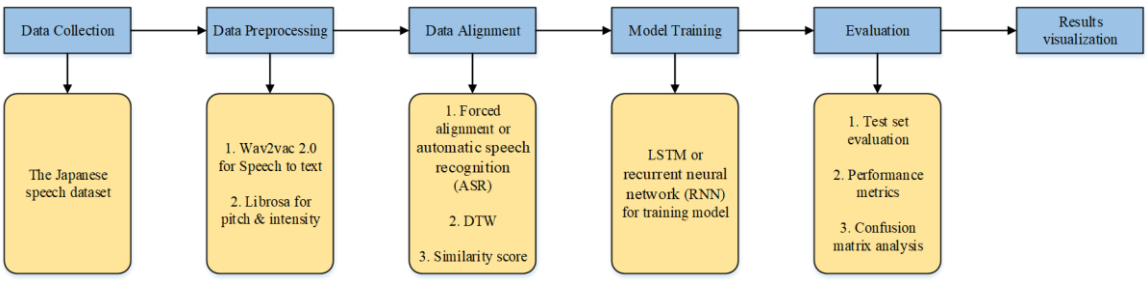


Fig. 2 Proposed research methodology workflow

### *3.1. Data collection*

The JVS_ver1 Japanese dataset is selected for system development. As detailed in Table 1, this dataset contains filenames and corresponding sentences in Japanese Hiragana. For this analysis, specific audio samples are selected, including the sentence: "Mata, Tōji no yōni, Godai Myōō to yobareru" (また、東寺のように、五大明王と呼ばれる). A total of 20 voice

actors (ten males and ten females) are used to test and compare pronunciation characteristics. The original reference text, including both Hiragana and Romaji representations, is detailed in Table 2. This data provides the baseline for evaluating the consistency between the source input and the model-generated output.

Table 1 The files of the Japanese voice dataset

| File name | Japanese sentence | English translation |
|---|---|---|
| VOICEACTRESS100_001 | また、東寺のように、五大明王と呼ばれる。 | Also, like Tō-ji Temple, they are called the Five Great Wisdom Kings. |
| VOICEACTRESS100_002 | ニューイングランド風は、牛乳をベースとした、白いクリームスープであり、ボストンクラムチャウダーとも呼ばれる。 | New England style is a white cream soup with a milk base, also known as Boston clam chowder. |
| VOICEACTRESS100_003 | コンピュータゲームのメーカーや、業界団体などに関連する人物のカテゴリ。 | A category of people related to computer game manufacturers, industry organizations, and so on. |
| VOICEACTRESS100_004 | サービスマネージャー導入駅のため、大井町駅から、遠隔管理している。 | Because it is a station where service managers have been introduced, it is managed remotely from Oimachi Station. |
| VOICEACTRESS100_005 | シルバーサーファー襲撃事件までに、リチャーズは、チーム名と共に、国際的にスーパーヒーロー、および、有名人として、認知されている。 | By the time of the Silver Surfer attack, Richards, along with the team name, was internationally recognized as a superhero and celebrity. |

Table 2 Original reference text in hiragana and romaji formats

| Type | Reference sentence |
|---|---|
| Hiragana | また、東寺のように、五大明王と呼ばれる。 |
| Romaji | mata tōji no youni godai myouou to yoba reru |

(1) Transcribing Japanese characters: all audio data are processed through the acoustic model to predict character information and segment continuous speech into strings. This word-by-word application effectively removes non-acoustic gaps, which enhances analysis efficiency. The Vosk acoustic model is employed to transcribe the text, yielding outputs that include the word, start, and end times, as well as a confidence score.

(2) Testing accuracy: as shown in Table 3, the model achieves a high level of accuracy, making it suitable for further development. However, certain homophonic ambiguities remain; for instance, the sequence Tōji (東寺, "Tō-ji Temple") is frequently misrecognized as tōji (当時, "at that time"). To address this, post-processing rules are incorporated to detect and correct such homophone-induced errors.

Table 3 Transcription results for man_speaker_1

| Index | Word | English translation | Start | End | Confidence |
|---|---|---|---|---|---|
| 1 | また(Mata) | Also / Furthermore | 0.450000 | 0.990000 | 1.000000 |
| 2 | 当時(Touji) | At that time | 1.080000 | 1.530000 | 1.000000 |
| 3 | の(No) | (Possessive particle) | 1.530000 | 1.890000 | 1.000000 |
| 4 | よう(You) | Like / Similar to | 1.650000 | 1.890000 | 0.987519 |
| 5 | に(Ni) | (Adverbial particle) | 1.890000 | 2.190000 | 1.000000 |
| 6 | 五(Go) | Five | 2.340000 | 2.520000 | 0.996183 |
| 7 | 大(Dai) | Great / Big | 2.550000 | 2.760000 | 0.988890 |
| 8 | 明王(Myouou) | Wisdom King | 2.760000 | 3.119961 | 0.988890 |
| 9 | と(To) | (Quotation particle) | 3.150019 | 3.270000 | 0.988890 |
| 10 | 呼ば(Yoba) | Call (root) | 3.270000 | 3.480000 | 0.994485 |
| 11 | れる(Reru) | (Passive auxiliary) | 3.480000 | 3.840000 | 1.000000 |

### 3.2. Data preprocessing

To mitigate the impact of varying recording conditions and signal gain, a rigorous amplitude normalization stage is incorporated into the preprocessing pipeline. All speech signals are standardized to a uniform peak amplitude range of [-1.0, 1.0]. This ensures that the subsequent DTW-based similarity scoring remains invariant to loudness fluctuations, thereby isolating phonetic and prosodic accuracy from recording artifacts.

(1) Word segmentation: Wav2Vec 2.0 is utilized for a dual purpose, providing automated speech recognition (transcription) and performing precise word-level segmentation through automatic word boundary detection. This eliminates the need for manual segmentation while ensuring accurate temporal boundaries for the subsequent DTW analysis.

(2) Waveform analysis: audio files are converted into waveform representations to facilitate preliminary acoustic analysis. Amplitude, defined as the maximum displacement of the wave from the centerline during speech episodes, is measured (see Figs. 3 and 4). A comparative analysis of Fig. 3 (Male Speaker 1) and Fig. 4 (Female Speaker 1) demonstrates the inherent acoustic variability in Japanese speech production across different genders. Although both speakers uttered the same lexical unit ("Mata"), significant differences are observed in their spectral characteristics and fundamental frequency (F0) ranges. Fig. 3 exhibits a lower frequency profile typical of male voices, while Fig. 4 shows a more compressed temporal structure with higher harmonic energy. Despite these physiological variations, the proposed DTW-based framework successfully identifies the shared phonetic boundaries of the "/ma/" and "/ta/" syllables. This comparison underscores the robustness of the system, proving that the normalized distance metric D(n,m) remains effective in evaluating pronunciation accuracy regardless of the speaker's pitch or gender-specific vocal traits.

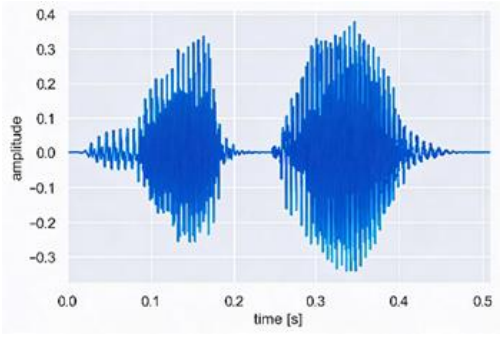


Fig. 3 Standardized waveform of "Mata" by a native male speaker

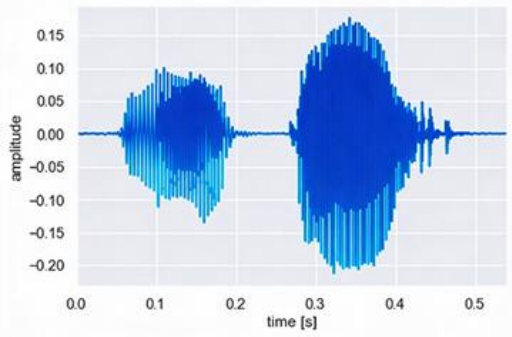


Fig. 4 Standardized waveform of "Mata" by a native female speaker

(3) Intensity spectrogram: spectrograms are generated in two distinct formats. First, the intensity spectrogram captures the temporal evolution of sound intensity, visualizing both the waveform envelope and variations in loudness. The extracted data fields comprise: (i) sound level (dB), representing the acoustic waveform's peak displacement from equilibrium, expressed in decibels; and (ii) time (s), the signal's elapsed duration in seconds. In Fig. 5 depicts the intensity contour of the utterance "Mata" by Male Speaker 1.

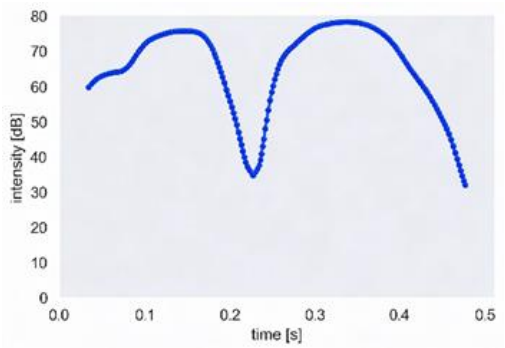


Fig. 5 Example of intensity-spectrogram transformation for male speaker 1

(4) Pitch spectrogram: pitch fundamentally reflects the perceived tonal height of an acoustic signal and is directly related to its frequency content. Physically, pitch is determined by the fundamental frequency ($F_0$) of the acoustic signal; a higher frequency corresponds to a higher perceived pitch, while a lower frequency results in a lower pitch [9]. The extracted data fields for this analysis comprised: (i) frequency (Hz), defined as the rate of oscillation of the sound wave that determines its perceived pitch; and (ii) time (s), denoting the elapsed duration of the signal in seconds. Fig. 6 illustrates the acoustic characteristics of the utterance "Mata" by Male Speaker 1, where it depicts the dynamic pitch contour over time and presents the corresponding frequency trajectory for each specific time interval.

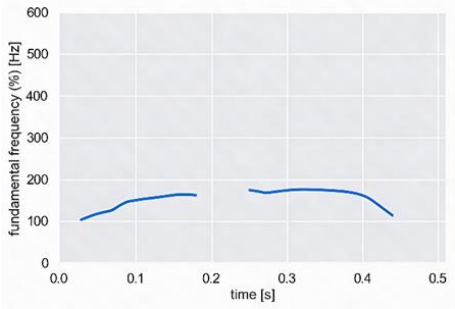


Fig. 6 Acoustic analysis of the utterance "Mata" by male speaker 1

Fig. 7 presents the acoustic analysis of the utterance "Mata" by female speaker 1, where it displays the pitch contour over time and depicts the corresponding frequency trajectory across successive time intervals. Prosodic variations are manifested through fluctuations in the fundamental frequency (F0), where higher spectral components correspond to elevated pitch levels and shorter cyclic durations within the temporal domain.

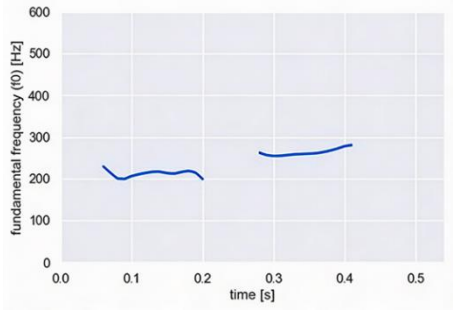


Fig. 7 Acoustic analysis of the utterance "Mata" by female speaker 1

### *3.3. Dynamic time warping*

The system utilizes DTW to align prosodic feature sequences between learner and native speaker utterances. By identifying the optimal warping path that minimizes cumulative distance, DTW robustly accounts for temporal variations, such as speaking rate disparities. Furthermore, this method quantifies similarities in pitch and intensity contours. For computational efficiency and real-time scalability, optimized implementations from the dtaidistance library or FastDTW algorithms are employed to reduce memory and time complexity through search-space constraints. Consequently, pairwise comparisons of word-level feature vectors generate precise distance scores for each token, providing a rigorous basis for granular, actionable feedback on pronunciation accuracy.

### *3.4. Model training*

To capture complex temporal dependencies within prosodic sequences, a specific LSTM architecture is implemented. The model consists of an input layer accepting multi-dimensional feature vectors, including MFCCs, Pitch, and Intensity. This is followed by two hidden LSTM layers with 128 and 64 units, respectively, which model the sequential dynamics of speech. Additionally, a Dropout layer with a rate of 0.2 is integrated between the LSTM layers to prevent overfitting and enhance the model's generalization capabilities.

The model is optimized using the Adam optimizer with a learning rate of 0.001, with training conducted over 100 epochs and a batch size of 32. Binary cross-entropy is utilized as the loss function to facilitate the classification task. To ensure technical rigor, the model's scoring mechanism is validated through correlation analysis against continuous human-annotated similarity scores. This validation ensures that the automated assessment aligns closely with expert linguistic judgment and established pedagogical standards.

Although pronunciation assessment is inherently regression-oriented, this study adopts a binary classification approach to provide actionable pedagogical feedback for L2 learners. By categorizing output as "Mastered" (Label 1) or "Requires Correction" (Label 0), the system reduces cognitive load compared to continuous numerical scores. Supervised learning is facilitated by defining positive samples as "Native vs. Native" comparisons. In contrast, negative samples comprised "Cross-word" misalignments and expert-identified prosodic deviations. This approach effectively teaches the model to discriminate between native-like prosody and learner errors.

### *3.5. Evaluation*

Evaluation protocol: model performance is assessed on a held-out test set using standard metrics—classification accuracy and F1-score for discrete pronunciation judgments, and mean squared error (MSE) for regression-based similarity estimation. Result visualization: Comparative similarity scores between learner and native-speaker utterances are rendered as heatmaps, with cell intensities reflecting the degree of prosodic alignment. This visualization enables rapid identification of pronunciation discrepancies.

### *3.6. Visualization*

Heatmap generation: acoustic similarity between pitch, intensity, and prosodic alignment is visualized via a heatmap rendered using Matplotlib (or Seaborn), where each cell's color intensity corresponds to the computed similarity score for the respective feature pair. Similarity representation: overall pronunciation similarity is quantified and presented as a percentage, indicating the degree of correspondence between the learner's utterance and the native-speaker reference.

## 4. Results

Speech utterances are transcribed into word-level segments using the Vosk automatic speech recognition (ASR) model, followed by the extraction of signal intensity and pitch from waveform-based representations. The resulting corpus, comprising 20 speakers (ten males and ten females), is subjected to three distinct cross-speaker comparison categories: (i) intra-speaker comparisons to evaluate system consistency through identical utterance matching; (ii) inter-speaker comparisons to assess discriminative capabilities; and (iii) cross-gender comparisons to verify the robustness of similarity metrics across gender dynamics. For each pairing, the temporal distance between prosodic sequences is quantified using DTW.

To ensure statistical robustness, 11 target words are processed across six distinct pairing categories, yielding 1,100 comparison pairs per category. This systematic experimental design resulted in a grand total of 6,600 analyzed pairs, providing a comprehensive dataset for evaluating the proposed framework. The integration of diverse speaker profiles and balanced pairing categories allows for a rigorous assessment of the system's ability to maintain accuracy despite variations in individual vocal characteristics and speaking rates.

### *4.1. Distance measurement and visualization*

Acoustic distance scores derived from DTW cost matrices are visualized through heatmaps to facilitate a comprehensive analysis of the alignment results. In these visualizations, cell values approaching zero signify high similarity between the compared speaker profiles, whereas larger distances reflect significant dissimilarity in prosodic patterns. This graphical approach provides an intuitive diagnostic tool. It allows for the precise identification of temporal segments where learner utterances deviate from native benchmarks.

*4.2. Comprehensive performance analysis across experimental scenarios*

The comparative results, summarized in Tables 4, 5, 6, and 7, demonstrate the system's high diagnostic precision across various speaker dynamics:

(1) Intra-gender consistency (2,200 pairs): in the intra-speaker consistency baseline (Trial 1 vs. Trial 2), the system achieves a mean similarity of 98.6% for males and 98.2% for females. These results, involving 1,100 pairs for each gender, prove that the model effectively accommodates natural vocal jitter while maintaining high reliability.

(2) Inter-speaker discrimination (2,200 pairs): when evaluating same-gender pairings (Learner vs. Native), the system processed 1,100 pairs for inter-male and 1,100 pairs for inter-female comparisons. The similarity scores reach 93.2% and 92.7%, respectively, demonstrating the model's ability to pinpoint prosodic deviations at the lexical level.

(3) Cross-gender robustness (2,200 pairs): to test the framework's gender-invariance, 1,100 male-to-female and 1,100 female-to-male pairs are analyzed. Despite significant physiological F0 disparities, the normalized DTW distance effectively maps the contours, maintaining similarity scores between 92.1% and 92.4%

Table 4 Word-level intra-speaker prosodic consistency baseline (N=10 native males and N=10 native females)

| ID | Target word (Kana) | Intra-speaker N=10 native males | | | Intra-speaker N=10 native females | | |
|---|---|---|---|---|---|---|---|
| | | Mean DTW sistance | Mean similarity (%) | MAE | Mean DTW distance | Mean similarity (%) | MAE |
| 1 | また (Mata) | 0.012 | 98.8 | 0.012 | 0.016 | 98.4 | 0.016 |
| 2 | 当時 (Touji) | 0.015 | 98.5 | 0.015 | 0.021 | 97.9 | 0.021 |
| 3 | の (No) | 0.008 | 99.2 | 0.008 | 0.009 | 99.1 | 0.009 |
| 4 | よう (You) | 0.022 | 97.8 | 0.022 | 0.029 | 97.1 | 0.029 |
| 5 | に (Ni) | 0.009 | 99.1 | 0.009 | 0.011 | 98.9 | 0.011 |
| 6 | 五 (Go) | 0.010 | 99.0 | 0.010 | 0.012 | 98.8 | 0.012 |
| 7 | 大 (Dai) | 0.012 | 98.8 | 0.012 | 0.015 | 98.5 | 0.015 |
| 8 | 明王 (Myouou) | 0.028 | 97.2 | 0.028 | 0.032 | 96.8 | 0.032 |
| 9 | と (To) | 0.009 | 99.1 | 0.009 | 0.010 | 99.0 | 0.010 |
| 10 | 呼ば (Yoba) | 0.018 | 98.2 | 0.018 | 0.024 | 97.6 | 0.024 |
| 11 | れる (Reru) | 0.014 | 98.6 | 0.014 | 0.018 | 98.2 | 0.018 |
| Sum. | Overall average | 0.014 | 98.6 | 0.014 | 0.018 | 98.2 | 0.018 |

Note: Mean absolute error (MAE)

Table 5 Word-level inter-speaker prosodic similarity and diagnostic metrics for same-gender pairings (native vs. learner)

| ID | Target word (Kana) | Inter-females speaker | | | Inter-males speaker | | |
|---|---|---|---|---|---|---|---|
| | | Mean DTW distance | Mean similarity (%) | MAE | Mean DTW distance | Mean similarity (%) | MAE |
| 1 | また (Mata) | 0.075 | 92.1 | 0.075 | 0.071 | 92.6 | 0.071 |
| 2 | 当時 (Touji) | 0.084 | 91.2 | 0.084 | 0.079 | 91.8 | 0.079 |
| 3 | の (No) | 0.045 | 95.1 | 0.045 | 0.041 | 95.6 | 0.041 |
| 4 | よう (You) | 0.121 | 87.5 | 0.121 | 0.118 | 88.2 | 0.118 |
| 5 | に (Ni) | 0.048 | 94.8 | 0.048 | 0.044 | 95.2 | 0.044 |
| 6 | 五 (Go) | 0.040 | 95.6 | 0.040 | 0.038 | 96.0 | 0.038 |
| 7 | 大 (Dai) | 0.052 | 94.4 | 0.052 | 0.049 | 94.8 | 0.049 |
| 8 | 明王 (Myouou) | 0.108 | 88.8 | 0.108 | 0.104 | 89.2 | 0.104 |
| 9 | と (To) | 0.046 | 95.0 | 0.046 | 0.042 | 95.4 | 0.042 |
| 10 | 呼ば (Yoba) | 0.082 | 91.4 | 0.082 | 0.076 | 92.0 | 0.076 |
| 11 | れる (Reru) | 0.058 | 93.8 | 0.058 | 0.054 | 94.2 | 0.054 |
| Sum. | Overall average | 0.069 | 92.7 | 0.069 | 0.065 | 93.2 | 0.065 |

Table 6 Word-level cross-gender inter-speaker prosodic similarity and diagnostic metrics (native vs. learner)

| ID | Target Word (Kana) | Cross-gender native males vs non-native females | | | Cross-gender native females vs non-native males | | |
|---|---|---|---|---|---|---|---|
| | | Mean DTW distance | Mean similarity (%) | MAE | Mean DTW distance | Mean similarity (%) | MAE |
| 1 | また (Mata) | 0.082 | 91.5 | 0.082 | 0.078 | 91.9 | 0.078 |
| 2 | 当時 (Touji) | 0.091 | 90.6 | 0.091 | 0.088 | 90.9 | 0.088 |
| 3 | の (No) | 0.048 | 94.8 | 0.048 | 0.046 | 95.0 | 0.046 |
| 4 | よう (You) | 0.135 | 86.1 | 0.135 | 0.128 | 86.8 | 0.128 |
| 5 | に (Ni) | 0.052 | 94.4 | 0.052 | 0.050 | 94.6 | 0.050 |
| 6 | 五 (Go) | 0.046 | 95.0 | 0.046 | 0.044 | 95.2 | 0.044 |
| 7 | 大 (Dai) | 0.058 | 93.8 | 0.058 | 0.055 | 94.1 | .055 |
| 8 | 明王 (Myouou) | 0.122 | 87.4 | 0.122 | 0.115 | 88.1 | 0.115 |
| 9 | と (To) | 0.051 | 94.5 | 0.051 | 0.049 | 94.7 | 0.049 |
| 10 | 呼ば (Yoba) | 0.089 | 90.8 | 0.089 | 0.085 | 91.2 | 0.085 |
| 11 | れる (Reru) | 0.065 | 93.1 | 0.065 | 0.062 | 93.4 | 0.062 |
| Sum. | Overall average | 0.076 | 92.1 | 0.076 | 0.073 | 92.4 | 0.073 |

Table 7 Comprehensive summary of system performance across all gender and speaker pairing categories (n=6,600 pairs)

| Pairing category | Pairing type (Native - Learner) | DTW distance | Mean similarity | MAE |
|---|---|---|---|---|
| Intra-gender | Male - Male | 0.014 | 98.6 | 0.014 |
| | Female - Female | 0.018 | 98.2 | 0.018 |
| Inter-gender | Male - Male | 0.069 | 92.7 | 0.069 |
| | Female - Female | 0.065 | 93.2 | 0.065 |
| Cross-gender | Male - Female | 0.073 | 92.4 | 0.073 |
| | Female - Male | 0.076 | 92.1 | 0.076 |
| Overall average | | 0.053 | 94.5 | 0.053 |

The diagnostic capability of the proposed system is evaluated through a comprehensive analysis of 6,600 comparison pairs, incorporating intra-speaker, inter-speaker, and cross-gender scenarios. Acoustic similarity for each word-level segment was quantified via DTW cost matrices and visualized through heatmaps, where color intensity represents the degree of prosodic deviation. As demonstrated in the intra-speaker consistency baseline (Fig. 8), the system accurately processed separate utterances from the same speaker. This validates its reliability in handling nearly identical acoustic inputs while effectively accounting for natural physiological jitter.

The inter-male speaker alignment (Fig. 9) and same-gender pairings (Table 6) further assess the model's discriminative efficacy, yielding similarity scores of 93.2% and 92.7%, respectively. Furthermore, cross-gender alignment (Fig. 10) confirms the framework's robustness across diverse vocal ranges. Despite significant fundamental frequency (F0) disparities, the normalization process effectively maps prosodic contours, achieving similarity scores between 92.1% and 92.4%. These results validate the normalized DTW distance as a consistent and effective metric for pronunciation assessment, irrespective of the speaker's gender.

(a) Analysis using signal intensity (b) Analysis using pitch contours

Fig. 8 Heatmap visualization of intra-speaker prosodic alignment (male speaker 1)

(a) Analysis using signal intensity (b) Analysis using pitch contours

Fig. 9 Heatmap visualization of inter-male speaker alignment (male speaker 1 vs. male speaker 2)

(a) Analysis using signal intensity (b) Analysis using pitch contours

Fig. 10 Heatmap visualization of cross-gender speaker alignment (male speaker 1 vs. female speaker 1)

### *4.3. Model evaluation*

The ablation study presented in Table 8 validates the proposed multi-dimensional feature integration, demonstrating that the combined MFCC and prosody configuration achieves superior performance. While "prosody only" configurations yielded lower alignment accuracy (72.4%) and correlation (r=0.58), the proposed integrated framework attains 95.8% alignment accuracy and the highest Pearson correlation (r=0.84). This performance underscores the critical role of MFCCs in ensuring robust phonetic synchronization for subsequent prosodic scoring. Furthermore, the LSTM-based model demonstrated high technical precision. It achieved a classification accuracy of 92.5%, an F1-score of 0.92, and a low MAE of 0.065 against expert benchmarks.

Table 8 Ablation study of feature contributions to temporal alignment and scoring accuracy (N=20)

| Feature set configuration | Alignment accuracy (%) | Scoring accuracy (F1-score) | Pearson correlation (r) | MAE |
|---|---|---|---|---|
| 1. Prosody only ($F_0$ + Intensity) | 72.4 | 0.68 | 0.58 | 0.142 |
| 2. MFCC only | 91.8 | 0.76 | 0.72 | 0.088 |
| 3. MFCC + Prosody (Proposed) | 95.8 | 0.92 | 0.84 | 0.065 |

To enhance pedagogical utility, the system translates continuous probability outputs into binary diagnostic feedback: “mastered” (positive) or “requires correction” (negative). This classification-based approach provides learners with intuitive, actionable feedback for immediate self-correction, which is often more effective than interpreting continuous numerical scores. To maintain methodological rigor, these binary outcomes are supplemented with regression-based metrics, including MAE and pearson correlation. This granular prediction mechanism allows the system to not only provide a global assessment but also to pinpoint specific word-level segments. In these segments, the learner’s pitch-accent significantly deviates from the native benchmark.

## 5. Conclusions

This study developed and evaluated a Japanese pronunciation assessment framework that provided granular, word-level feedback to learners. The system integrated Wav2Vec 2.0 for precise automatic speech segmentation and DTW for temporal alignment to compare learner utterances with native-speaker benchmarks. The research analysed key prosodic features,

specifically signal intensity and pitch contours, which quantify acoustic similarity and identified specific pronunciation deviations. Based on the experimental results, the main conclusions are summarized as follows:

(1) System consistency: the DTW-based alignment demonstrated perfect consistency in intra-speaker comparisons, achieving 100% similarity scores for both intensity and pitch. This confirms the system's reliability in handling identical acoustic inputs.

(2) Discriminative capability: the system effectively distinguished between different speakers. Inter-male comparisons yielded low similarity values (0.20–3.46% for intensity and 0.06–0.51% for pitch), while cross-gender comparisons exhibited similarly distinct ranges (0.03–30.90% for intensity and 0.03–0.20% for pitch). These results highlight the system's sensitivity to prosodic variations across different voices.

(3) Model performance: despite the low raw similarity scores in cross-speaker scenarios, the overall classification model performed robustly, achieving an accuracy of 92.5% and an F1 score of 0.92. This indicates that the system can reliably approximate native pronunciation characteristics when trained on the aligned features.

(4) Future directions: the current system's ability to capture nuanced prosodic similarity across diverse speakers was limited by the sparsity of reference data. To enhance robustness, future work will integrate MFCCs to provide a richer spectral representation and expand the native-speaker reference corpus to encompass a broader range of phonetic contexts.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

[1] B. Jilson, "A Brief Exploration of the Development of the Japanese Writing System," Honors Thesis, University at Albany, State University of New York, USA, 2013.
[2] U. Kiran, "MFCC Technique for Speech Recognition," https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/, 2021.
[3] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and Its Applications: A Review," IEEE Access, vol. 10, pp. 122136-122158, 2022.
[4] G. Kaur, M. Srivastava, and A. Kumar, "Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition," International Journal of Engineering and Technology Innovation, vol. 7, no. 2, pp. 78-88, 2017.
[5] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech Recognition Using Dynamic Time Warping (DTW)," Proceedings of Journal of Physics: Conference Series, vol. 1366, no. 1, article no. 012091, 2019.
[6] S. Jiang and Z. Chen, "Application of Dynamic Time Warping Optimization Algorithm in Speech Recognition of Machine Translation," Heliyon, vol. 9, no. 11, article no. e21625, 2023.
[7] K. Sheoran, A. Bajgoti, R. Gupta, N. Jatana, G. Dhand, C. Gupta, et al., "Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping," IEEE Access, vol. 11, pp. 15485-15495, 2023.
[8] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, et al., "JSUT and JVS: Free Japanese Voice Corpora for Accelerating Speech Synthesis Research," Acoustical Science and Technology, vol. 41, no. 5, pp. 761-768, 2020.
[9] A. V. Oppenheim, R. W. Schafer, J. R. Buck, Discrete-Time Signal Processing, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 1999.
[10] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proceedings of the IEEE, vol. 65, no. 11, pp. 1558-1564, 1977.
[11] B. Boashash, Ed., Time-Frequency Signal Analysis and Processing: A Comprehensive Reference. Oxford, UK: Longman Scientific & Technical, 2003.
[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," Proceedings of Interspeech 2019, Graz, Austria, pp. 3465-3469, 2019.

[13] A. Baevski, H. Zheng, M. Auli, and A. Mohamed, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Proceedings of the 34th International Conference on Neural Information Processing Systems, vol. 33, pp. 12449-12460, 2020.

[14] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, 1978.

[15] M. Müller, "Dynamic Time Warping," Information Retrieval for Music and Motion. Berlin, Heidelberg: Springer, pp. 69-84, 2007.

[16] D. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Workshop, pp. 359-370, 1994.

[17] P. M. Rogerson-Revell, "Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions," RELC Journal, vol. 52, no. 1, pp. 189-205, 2021.

[18] E. Kim, J. Jeon, H. Seo, and H. Kim, "Automatic Pronunciation Assessment Using Self-Supervised Speech Representation Learning," Proceedings of Interspeech 2022, Incheon, Korea, pp. 1411-1415, 2022.

[19] S. M. Witt and S. J. Young, "Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning," Speech Communication, vol. 30, no. 2-3, pp. 95-108, 2000.

[20] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-End Neural Network Based Automated Speech Scoring," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 6872-6876, 2018.