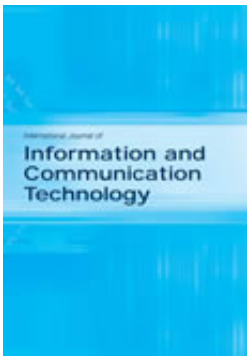


International Journal of Information and Communication Technology

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

Development of an AI-assisted spoken language assessment system for Japanese language teaching

Li Zhang

DOI: 10.1504/IJICT.2026.10076058

Article History:

Received:	08 October 2025
Last revised:	19 November 2025
Accepted:	02 December 2025
Published online:	13 February 2026

Development of an AI-assisted spoken language assessment system for Japanese language teaching

Li Zhang

School of Foreign Languages,
Nanyang Institute of Technology,
Nanyang, Henan, 473000, China
Email: 19837722586@163.com

Abstract: This project develops an AI-assisted spoken language assessment system to enhance Japanese language instruction. By integrating voice recognition, pronunciation analysis, and fluency scoring, the system provides reliable evaluations comparable to those of professional ratters. Utilising multilingual datasets and data augmentation, it reduces language learning anxiety and improves recognition accuracy. Since language anxiety negatively affects second-language acquisition – particularly in oral proficiency – this study aims to support Japanese learners and increase their speaking confidence. While prior research has demonstrated the benefits of ICT tools and flipped classrooms for pronunciation self-monitoring, limited studies have applied AI for comprehensive oral evaluation. The proposed four-step methodology includes data collection, feature extraction, model development, and validation. Informed by sentiment analysis and multilingual corpora, the system achieved an accuracy of 97.3% using a two-stream LSTM model, while translation-based augmentation improved Japanese sentiment analysis accuracy by 6.58%.

Keywords: AI-assisted assessment; spoken language evaluation; Japanese language teaching; speech recognition; natural language processing; NLP; automated scoring; pronunciation analysis.

Reference to this paper should be made as follows: Zhang, L. (2026) 'Development of an AI-assisted spoken language assessment system for Japanese language teaching', *Int. J. Information and Communication Technology*, Vol. 27, No. 10, pp.90–111.

Biographical notes: Li Zhang holds a PhD and serves as a Lecturer. Her primary research area is foreign languages and literatures.

1 Introduction

It has been found that people who think a lot about language also have trouble learning it. Since he began learning English, the student in this study has been under a lot of stress (Gregersen, 2020). When he was five years old, he started taking English classes. He had to go to English school every day when he was six years old. Besides that, he joined two groups where kids from various places could spend time together. One was in Australia and the other was in the US. Though he has been a teacher for thirteen years, he is often in a bad mood. It is the worst when he has to talk. There are two different goals for this

project. The student was scared to learn a new language. Why was that? That is the study's central question. The second goal is to learn how he handled his fears. A lot of research has shown that stress makes it harder to learn a language. But not as much research has looked at how people who are learning a language deal with stress. This way of looking at an event that lasts for a long time has only been used in a few studies (Ebbinghaus, 2020). We want to fix these things with this work. This term refers to the fear-based emotions, such as worry, that come up when you learn or use a second language. Worry is the thing that SLA experts have looked into the most because it is so strong and common. What was the worst thing that could happen to English learners? It was bad for speaking, according to a study on the subject. It has been found that one of the things that scares people the most is having to talk in public.

People were asked to rate how they felt when they used a second language in a study. 85% of those people said they were scared. There is also a strong link between how anxious students are about learning English and how well they speak English, as shown by Said's study. But it has been found that people from different countries deal with worry in very different ways. It can be scary not to be able to 'follow the rules' in places with strict rules. People should not stand out in Japan, where this study took place. That's how they live there. This is why people who are learning the language there get scared when they try to say it (Sano Nakao and Reinders, 2022). Some Japanese language learners are stressed because they do not do well in school, they have doubts about themselves, and they cannot think straight. Japanese students were asked to name three English-related things. As technology keeps getting better, there are a lot of new ways to teach. One of these ideas is the flipped classroom (FC) method. It might help teachers and students get along better, and it might also make students want to learn Japanese more, which would make Japanese education better overall (Shrestha et al., 2020). The study's goal is to build a theoretical foundation for changes in how Japanese is taught so that Japanese education works better.

A look into whether or not FC can be used in Japanese training and how it can be used in Japanese teaching will be conducted. FC lets students watch short videos before class, and they can change how long the video plays and how it moves forward based on how they learn best. They can skip over parts they already know, and they can watch it more than once to get a better grasp on it. Teachers need to remember that when they make micro-videos, they should think about how hard the information is. This will keep students interested in learning.

Japanese and give teachers and students more time to talk and learn on their own, which is suitable for students' learning (Chhabra and Singh, 2020). They should make videos with simple, easy-to-understand knowledge and teach more difficult things in school. Teachers need to remember this because they need to know how hard the work is. Finding a way to measure how well task-based and FC teaching methods help teach Japanese is the primary goal of this study. Another goal is to improve the usefulness of teaching and Japanese skills in both of these methods. These evaluation methods have been used in the past to see how well task-based and FC Japanese language teaching works. They are the correlation feature extraction method, the particle swarm optimisation evaluation algorithm, and the multi-source information resource service method. A hybrid evaluation method that uses both online and offline parts to find out how well training is working could work. But there is a problem: it takes a lot of money

to study how teaching Japanese in FC and task-based settings changes students. There is a model in the literature that matches the way people move in Taekwondo.

This way, teachers can correct their students' movements, give them sports tips, and let them practice by simulating moves. However, this approach is not very good at self-adaptive optimisation when it comes to checking how well the FC and task-based teaching mode work for teaching Japanese (Yu and Liu, 2023). Japan and China have talked to each other for a long time, even though they are from different countries. In China, this is making more and more people want to learn Japanese. There were 3.985 million people in the world, but 1.04 million were learning Japanese as a second or foreign language (JSL/JFL) in that one country (Wang and Zheng, 2021). From 2009 to now, this is a rise of 26.5%. Most people in mainland China speak Japanese because it is beneficial for business, and people from other countries use Japanese in their own work, like in Japanese cartoons. These two things likely are what caused this trend. We need to know how JSL/JFL teachers in China feel about this because more and more people want to learn Japanese. It is critical to pay attention to grammar when you are learning Japanese. In order to use different tenses when drawing lines, people who are learning Japanese as a second language have to change the way they put words together. This is shown by the fact that Japanese verbs and nouns must be conjugated with different tenses, but words do not (Maie and Godfroid, 2022).

If you are learning Japanese as a second or foreign language, verb conjugations are the most essential part of your lessons. This is because they are the building blocks of Japanese syntax. When you conjugate a word, you change it from its basic parts to forms that come from them. You can say many things with this. Japanese books teach three simple ways to change the form of a verb in Chinese. The tenses shown here are the same in English, but they are written in two different ways. That is, the simple past tense (ta), the simple future tense (simple), and the present progressive tense (simple and simple future tense) are all the same.

The paper is organised into five sections to provide a coherent and systematic analysis of the research topic. Section 2 presents the literature review on the development of an AI-assisted spoken language assessment system. Section 3 outlines the proposed methodology for Japanese language teaching. Section 4 discusses the results, and Section 5 concludes the study.

1.1 The novelty of this work

Three main features of this work make it novel:

1 Translation-based data augmentation specifically designed for Japanese sentiment analysis, which achieved a 6.58% accuracy improvement through cross-lingual transfer learning from multilingual Amazon review corpora.

2 Language model replacement (LMR) technique, which adapts pre-trained end-to-end ASR models to the Japanese language domain without requiring full model retraining, thereby reducing computational costs while maintaining high recognition accuracy.

3 Two-stream LSTM architecture for dynamic Japanese sign language (JSL) recognition integrated with spoken language assessment, which achieved 97.3%

accuracy through hierarchical feature extraction (distance, angle, direction, and variation) across segmented video frames through segmented video frames.

Our method is ideally suited for holistic Japanese language training since it offers thorough evaluation across numerous dimensions, including pronunciation, fluency, sentiment, and gesture detection, in contrast to previous systems that concentrate on isolated components (pronunciation alone or fluency only).

1.2 Contribution of the study

By creating an AI-assisted spoken language assessment system specifically designed for Japanese language instruction, this study advances the expanding field of technology-enhanced language learning. The suggested approach combines voice recognition, pronunciation analysis, and fluency scoring to produce unbiased and scalable evaluations, in contrast to conventional assessment techniques that primarily rely on human raters. Through the utilisation of multilingual sentiment analysis datasets, novel feature extraction approaches, and LMR methods, the research improves the precision and dependability of competence evaluations in a variety of learner scenarios. This method addresses computing efficiency through sophisticated machine learning and deep learning architectures, in addition to increasing the accuracy of language evaluation. Additionally, by connecting theoretical understandings of second language acquisition with real-world applications in online learning environments, the research improves pedagogical practice. By giving regular feedback, lowering rater subjectivity, and encouraging student autonomy through self-monitoring, the method offers instructors invaluable support. Furthermore, the results demonstrate how adaptive modelling, multimodal feature alignment, and data augmentation may be used to get around issues like small datasets and language-specific complexity. When taken as a whole, this research advances AI-driven language evaluation techniques and advances the reform of Japanese language instruction in a more technologically mediated and globally interconnected educational environment.

2 Literature review

2.1 Special Mora education in Japanese

This is how you put together a single Japanese word: a Mora. One word, on the other hand, has two parts: a regular part and a unique part. These things happen when a particular kind of Mora is used. Japanese trainers who do not speak Japanese as their first language have a hard time telling the difference between lesson units that teach different kinds of rules. Japanese speakers can mean the difference between standard and strange behaviour with just one word. A lot of people who work in special needs education (Vendityaningtyas et al., 2020) talk about how hard it is to tell when someone has special needs. The most important study on this subject says that students should read the rules before they speak. People learning the special mores need to know how long the sounds last before they can understand the correct way to say them. Students will not be able to get to the right amount of the Mora if they can not tell the difference between how different mores say the same Mora. Kids were also told how important it is to do things

that help them see how much they've grown. They should be able to figure out what's wrong and fix it on their own since they know some applicable rules about how sounds work in Japanese. This study looked at Korean trainees and how the way Korean syllables are put together helped them learn some social rules. English was taught to the kids at stages above and below. These people did this work.

2.2 The use of ICT to support self-monitoring of pronunciation learning

In the past few years, Japanese speech training has paid more attention to self-monitoring for sound recognition (Matsuzaki, 2012). A study that taught speech with magnetic tracking and a learning management system (LMS) is one of the things that is talked about. People could take lessons whenever they needed to because they were given in person and online. As a test, some students who were learning Japanese as a second language (JSL) were allowed to join the class for one term. They used on-demand independent learning tools to help them learn more, as shown by the term papers they turned in for the final study. Prosody Tuner was used to test speech as part of a bigger project that looked at technology for information and conversation. To make sure that every word was heard at the same time, the way that a pupil's and a model's voice spoke was broken down into consonants and vowels for that study project. The learner's voice was in the middle of the screen, and the model's voice was at the top. This was done so it would be easy to play back and so people could check out both the sound and the picture at the same time. About two-fifths of users said the software helped them understand how words were pronounced. A speech analysis tool helped them improve their pronunciation by letting them listen to and check their own pronunciation (Lee et al., 2022). Some people said the software was good at understanding words. They also learned how to change how they behaved by being taught how to watch themselves. You can understand them better if you listen to them talk and compare it to how most Japanese people say words. This made the speech better by changing it. Keep an eye on yourself with information and communication technology (ICT). This has been shown to help you speak Japanese better.

2.3 Informal learning in Japanese English education

There is a government in Japan that makes sure schools teach English. We spend a lot of time reading and writing in this class because it helps them prepare for tests that schools use to decide who to let in (Crane and Sosulski, 2020). A lot of college English language programs test students a lot while they teach them how to talk and listen. Intercultural competence is the ability to speak to people from different cultures without much trouble. Not much study has been done on how it affects language learning this way. This shows how important it is to learn in both formal and informal situations if you want to understand other cultures better (ICC). Japan sees a lot of TOEIC scores to see how well English learners are doing, but not as much in foreign skills. People from all walks of life and income levels learn from each other all the time. This is called 'informal learning'. This is another way to learn besides what you learn in school. If you want to learn English at university, you do not learn enough about it or get the most out of relaxed tasks. Schools in Japan use grammar translation to help kids learn how to read and understand what they read (Deacon and Miles, 2023). This is how English is taught in

Japan. That's good, because it will help you read and write better on tests, but it might not help you in real life. These skills will help you understand what you read better.

It's a shame that Japanese students' English has been getting worse over the years. A new poll shows that people who do not speak English no longer exist in 92 of the 116 countries that were polled. There has never been a number this low. Another study found that Japanese students did not want to learn English, so this makes sense. Japan might need to change how English is taught so that students can speak English better with each other. Talking openly with people from different countries could be one way to do this. When people really take part, they show what's called 'interactional competence' (Nagashima, 2025).

2.4 Information seeking and media usage

Disasters make people want to know more about what's going on so they can make better decisions (Kawasaki et al., 2013). Every time they get more and better information, they learn more about risk and act better on what they know. There are many types of media and information sites that can be used to find information. A lot of the time, people use TV, radio, newspapers, and social networks to find out what's going on during significant crises (Kawasaki et al., 2018). The best means will depend on when the accident happened, what kind of event it was, and who was there (Lachlan et al., 2014). Several studies have also shown that some types of media are not as helpful as they may seem. Most TV stories are short and have a lot of pictures. They make people feel something. But written media are longer and go into more detail, so they might do a better job of spreading facts. Human networks, on the other hand, help confirm information among ties that are thought to be reliable.

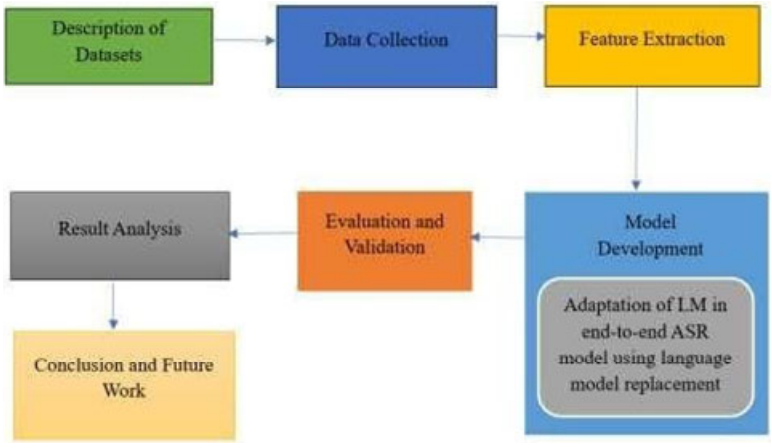
The limitations of static and unidimensional analysis in English teaching quality assessment are addressed in this study (Yan, 2025) by proposing a dynamic assessment method based on multimodal cognitive transfer modelling. By simultaneously gathering four-dimensional data from the teaching scene (speech, vision, text, and physiological signals), we can build a two-channel LSTM-cognitive state space model. In the knowledge transfer channel, we can quantify students' cognitive state transfer trajectories using the ACT-R cognitive architecture. In the teaching intervention channel, we can model the feedback mechanism of teachers' strategy adjustment using a dynamic causal map. This paper (Kong, 2025) suggests an AI-augmented POA framework to fix major problems with the production-oriented approach (POA) to teaching English as a second language, such as long wait times for feedback, ineffective development of contextual tasks, and inadequate allocation of resources. In order to improve POA's 'drive-facilitate-evaluate' closed loop, we created a two-engine design that incorporates dynamic task generation, multimodal resource recommendation, and multidimensional evaluation.

Incorporating AI technology and a voice knowledge recognition algorithm, this work (Li, 2024) suggests a sophisticated scoring mechanism for spoken English self-learning systems. Students' vocal expression skills will be better assessed and they will receive more tailored and focused learning assistance through the use of intelligent technology in this new mechanism.

3 Methodology

The detailed study process is depicted in Figure 1, which starts with dataset description and collection and moves on to feature extraction, model building with language model adaptation, evaluation, result analysis, and future work.

Figure 1 The suggested AI-assisted spoken language evaluation system's framework (see online version for colours)



3.1 Description of datasets

This study used review files from online stores to make mood analysis models that could add to the data. You can find the Multilingual Amazon Reviews Corpus on Amazon Web Services (AWS). It told us what we already knew. People talked about books on Amazon in English, Japanese, German, French, Spanish, and Chinese from 1 November 2015 to 1 November 2019. People in the US, Japan, Germany, France, Spain, and China sent their thoughts. The dataset (Alkhushayni and Lee, 2025) is made up of many parts, such as the reviewer ID, review text, review title, rating (from 1 to 5), product ID, and product category. There are many reviews for many things out there. The reviews picked for this study are all about 'beauty' items. It's best to choose a few things because each type of customer review is different. In a tech report, it might say, "this is very quick and easy to setup." You might also want to add the line "they are very comfortable when worn" to the report. When we only sell one type of item, people can tell us about their own thoughts and ideas. There is no connection between Japanese and any other language, so it is not like them. Some people think it is not English. Japan has three ways to write. Latin, on the other hand, is always used the same way.

The words in these three languages are not all in the same order. The rules of English say that words should go from 'what' to 'what'. The subject, verb, and object are more like the parts of a word in English. These languages also have very different word names and rules for how to talk and act. There are three groups in this image that show the number of good and bad reviews. This is shown in Figure 2. There is a slight difference between the groups of plus and minus numbers, even though all three numbers are spread out. List of Table 1's good and bad review counts for all three groups. This picture goes

with Table 1. It will be hard for the mood analysis tool to find the necessary parts because the files are too small. If you act this way, it will be hard for the model to figure out what is best and most true.

Figure 2 Bar plot of sentiment (see online version for colours)

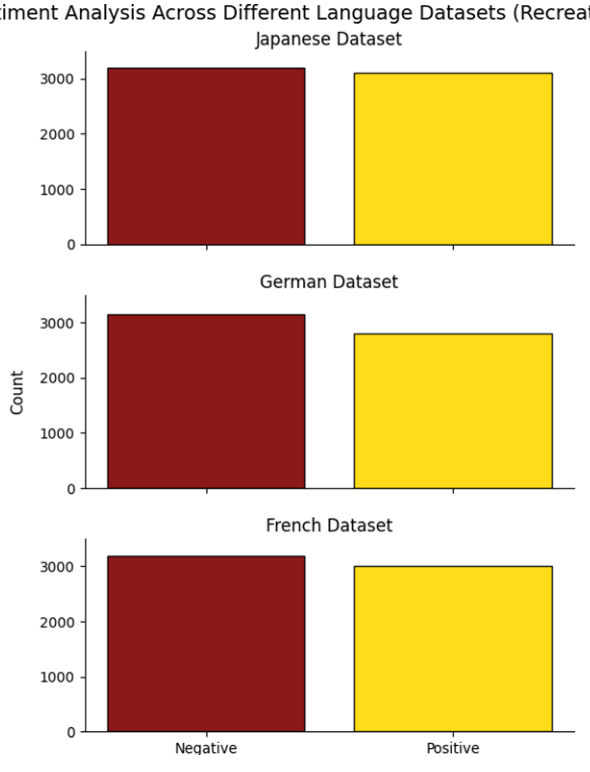


Table 1 How many positive and negative samples are there in each of the three databases

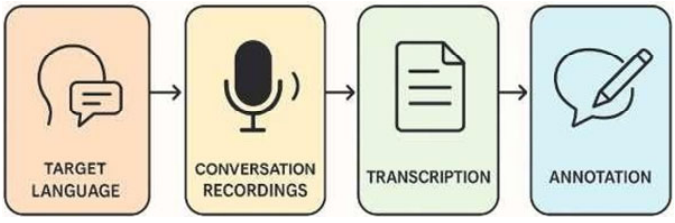
Set of data	*Number of negative samples*	*Number of positive samples*
French	2,961	3,650
English	2,679	3,205
Japanese	3,281	3,462

3.2 Data collection

The study was done in two places between 15 December 2022 and 15 January 2023. MS Forms was used to make the website link that goes to the online poll form. The snowball sample method was used to find people who wanted to take part (Tran et al., 2023). At first, the study only asked a small group of students who were taking more than one class and were serious about or interested in Japanese studies to take part. It was emphasised that everyone in the class should get the poll's URL and then share it with their peers. People who filled out the survey knew what the poll was about before they started. It was written on the first page. These people did not feel pushed to take part in this study.

Everyone who took part in the survey knew everything there was to know about the privacy policy before they started. It is important to note that no personal information was recorded about the people who took part. People who filled out the poll did not get anything in return, and they could quit at any time (Yi et al., 2023). It was okay with the Ethics Committee of Tokushima University's Graduate School of Science and Technology to do this work. A four-stage linear pipeline for processing linguistic data is shown in Figure 3. Raw spoken input in the target language is the first step in the process, and it is then recorded as conversation recordings. Transcription is used to turn the audio into written text. The last step is annotation, which includes classifying and labelling the text to improve its usefulness for machine learning or linguistic analysis applications.

Figure 3 The four-step language data collection, transcription, and annotation process (see online version for colours)



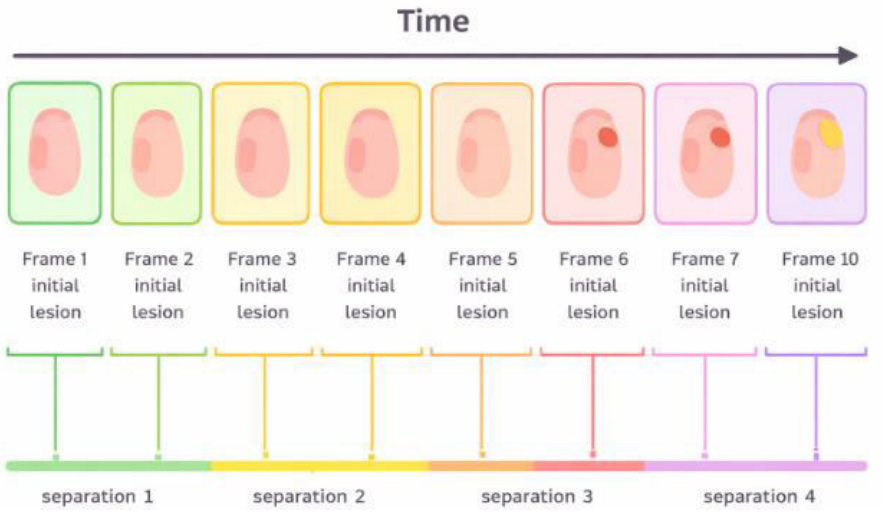
3.3 Feature extraction

The dataset was split into four parts so that information about long-term behaviours could be kept separate from information about what signs meant. The type of rotation, the distance from the palm, the angles made by the fingers and finger joints, and the distance and angles from the tip of the thumb helped us figure out what each division was made of. The point of these features is to show what makes each finger sign unique. This part (Kakizaki et al., 2024) talks about four types of traits that were discussed in the last part. We used a technique called 'moving average calculation' to deal with the fact that the process of feature extraction used a number of different picture frames. When the finger spelling changed quickly, this was a must. This method makes sure that there are always the same number of features, no matter what frames are used to show each figure. The value of each trait was added up across all four parts of each frame (see Figure 4). Our plan worked because of this. This study was told to use this 'averaging' method on three of the four traits. These four parts were found in all frames based on the numbers that were given. Figure 4 shows that the process of segmentation is complete when there are twelve frames.

There were 3,529 features to choose from, such as 760 lengths, 2,520 angles, 60 orientations, and 189 types. That is, there were 3,529 traits altogether. But you could pick and choose which traits to use, and it turned out that the models did not need all of them to be very accurate. To get feature values, hand coordinate data from the RGB picture data is used after the data has been collected. There are 21 hand positions, and MediaPipe is used to measure them in X, Y, and Z. Things like the picture size and the distance between the hand and the camera are kept similar so they do not make a big difference. This work suggests four different types of feature values. This is what Matsuoka came up

with when he used MediaPipe and SVM to try to learn American sign language. You can measure these things with ‘distance’ and ‘angle’.

Figure 4 To determine feature distance, angle, and finger direction, divide the video frames into groups (see online version for colours)



3.4 Two-stream LSTM model configuration and training details

The two-stream LSTM architecture processes both spatial and temporal features through parallel streams. Table 2 provides comprehensive details of model hyperparameters, acoustic feature specifications, and ASR training configuration used in our implementation.

Table 2 Detailed model parameters and configuration

Category	*Parameter/feature*	*Specification*
Two-stream LSTM hyperparameters	Batch size	32
	Optimiser	Adam
	Learning rate (LR)	0.001
	Epochs	100
	Early stopping patience	10 epochs
	LSTM hidden units	128 per stream
	Dropout rate	0.3
	Input shape – stream 1 (spatial)	(2, 835)
	Input shape – stream 2 (temporal)	(2, 63)
Acoustic features	Mel-frequency cepstral coefficients (MFCC)	13 coefficients + Δ + ΔΔ = 39 features
	FBank (filter bank)	80-dimensional log mel-filter bank energies
	Prosody features	Pitch (F0), energy, duration statistics

Table 2 Detailed model parameters and configuration (continued)

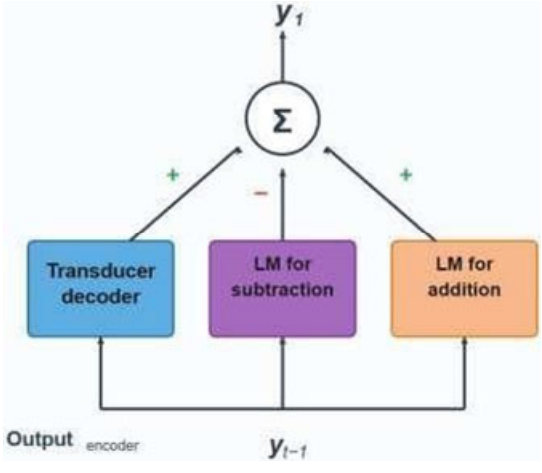
Category	*Parameter/feature*	*Specification*
ASR training loss	Loss function type	Hybrid (CTC + cross-entropy)
	CTC loss weight	0.3
	Cross-entropy (CE) loss weight	0.7
	Total loss formula	L_total = 0.3 × L_CTC + 0.7 × L_CE
	Encoder-decoder architecture	End-to-end attention-based encoder-decoder

The hybrid loss function combines CTC for alignment-free sequence learning with cross-entropy for character-level prediction, enabling robust Japanese speech recognition without explicit phoneme-level alignment.

3.5 Adaptation of LM in an end-to-end ASR model using LMR

The language model in a regular E2E automatic speech recognition model that has already been trained should be changed to make speech recognition better in the target area (Mori et al., 2024). To do this, first, a rough idea of the 'implicit language information' that is already stored in the ASR model that has already been trained is made. For the inference step, this language information is used to get rid of the language information that was learned from the source topic data that the ASR model was trained on in the first place. Then, language data from the target domain that was taken from the LM of an ASR model that was trained independently is mixed with language data that is already in the modified ASR model. This is done in a way that is similar to Shallow Fusion. Figure 5 is a picture that displays the suggested method.

Figure 5 Identifying the target domain using LMR for Japanese speakers (see online version for colours)



If there is no language model merging method, the ASR model thinks that the following things will happen:

$$\hat{y} = \arg\max_{y} \{\log P_{source}(y / x)\} \tag{1}$$

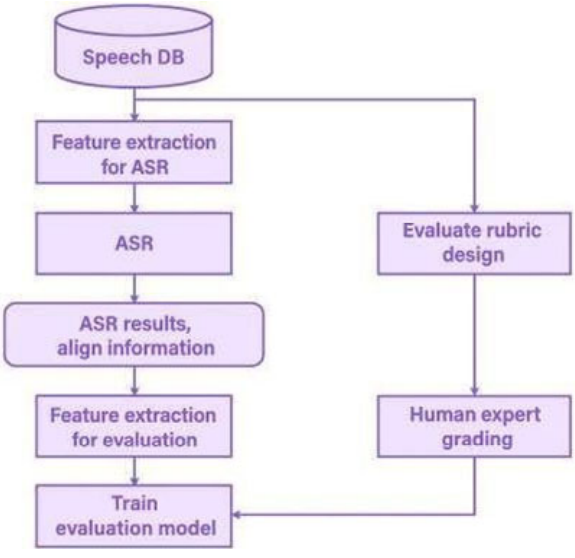
3.5.1 Human rating criteria and rubrics

Five expert Japanese instructors (10+ years teaching experience) independently rated each speech sample using this five-point Likert scale rubric. Inter-rater reliability was assessed using Cronbach's alpha ($\alpha = 0.82$), indicating strong agreement among human raters.

Table 3 Fluency and pronunciation scoring rubric

Criterion	*Score 1 (poor)*	*Score 3 (average)*	*Score 5 (excellent)*
Pronunciation	Multiple mispronunciations; special Moras incorrect	Some errors in special Moras; generally intelligible	Near-native pronunciation; accurate special Moras
Fluency	Frequent pauses (>3 sec); hesitant speech	Occasional pauses; moderate flow	Smooth delivery; natural rhythm and intonation
Intonation	Monotone; incorrect pitch patterns	Some pitch variation; acceptable patterns	Natural Japanese pitch accent; appropriate prosody
Speaking rate	Too slow (<2 Mora/sec) or too fast (>7 Mora/sec)	Moderate pace (3–5 Mora/sec)	Natural pace (4–6 Mora/sec)

Figure 6 Training proficiency evaluation model diagram (see online version for colours)



3.6 Evaluation and validation

Parts of fluency that can be used to test proficiency in more than one area can be used to teach a model how to judge proficiency automatically. After that, you could let the model

decide how well you speak (Denga and Denga, 2024). The model learns how to judge success all the way through, as shown in Figure 6. This method has a special test that people who do not speak the language as their first language use to figure out different sound features in speech. Some of these sound traits are segmental features, intonation, and pace. Voice notes are typed up by a machine that can automatically recognise speech (ASR). The next step is to use a forced-alignment method to find the right word and phoneme pairs at the right time. This is what makes these flow traits possible. There are also times when phonemes and words happen, along with the sound numbers that go with each set. Each word and phrase can be used in several different ways to find out how fast you speak. Then, raters who are very good at showing language traits and scores are used to teach the models that are used to rate ability. The people on the review team are English professors whose main job is to teach English, and English teachers who have taught before. To help them get used to the score rubrics, the raters were trained ahead of time. They did this to make sure each number was the same.

4 Result

The study's results are shown here. The study checked how well mood analysis worked when three types of translation data were added. The scores are given here. To find out how well machine learning models work, it is essential to test and compare them in the area of natural language processing (NLP). A lot of the time, assessment tools are used. Some of these are memory, accuracy, F-score, and precision. There is a number that tells you how well the model can put variables into the right groups. The tests were used to see how well the machine learning models did when given more data. For every tongue, this was done. There is a table called a confusion matrix that can be used to measure how well classification is going. It shows how different predicted and real classes are. This table can be used to find the F-score, accuracy, precision, and memory. Table 2 shows how to do this.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F\text{-}score = \frac{2 \times precision \times recall}{precision \times recall} \tag{5}$$

Table 4 Matrix of confusion

	Thoughts: no	*Guessed: yes*
Truth: no	Real negative	Not true positive
Real: indeed	Not true negative	Real and true

Equation (1) shows that it is right when it gets things right about half of the time. Accuracy is like a number that tells you how often the filter is right when it says 'yes'. Equation (3) tells us how often the guess is correct when it says 'yes'. This is where the word 'recall' comes from. To find out how good a binary classification is, you can look at the F-score [equation (4)] along with the accuracy and memory numbers. The data augmentation method was used on three sets of data written in three different languages to look for emotion. This session talks about what happened. There are graphs in Figure 7 and Figure 7 that show how well Google Translate and Deep Translate add data. This method can help make mood studies more accurate across a number of language sets. You can use Google Translate with it. When the data extension method was used, the model did better with both the French and Japanese datasets. Just in the Japanese group, there was a significant rise of 6.58%. This method did not make mood analysis work better in the German group, though.

Figure 7 Google Translate's accuracy usage graph for data augmentation (see online version for colours)

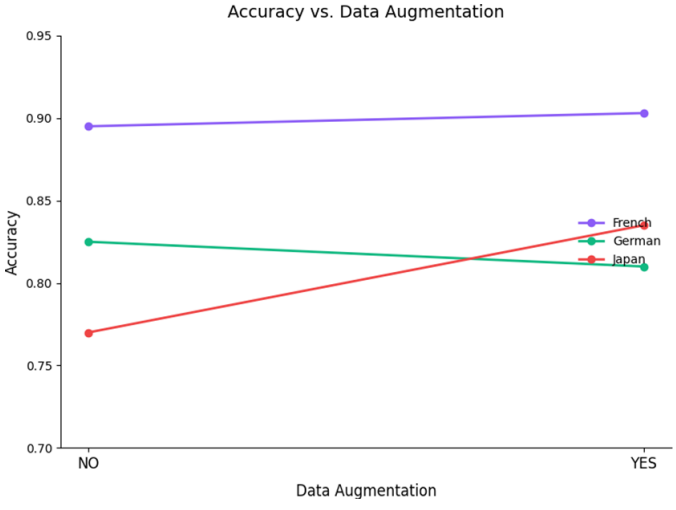


Table 5 Japanese language proficiency as perceived

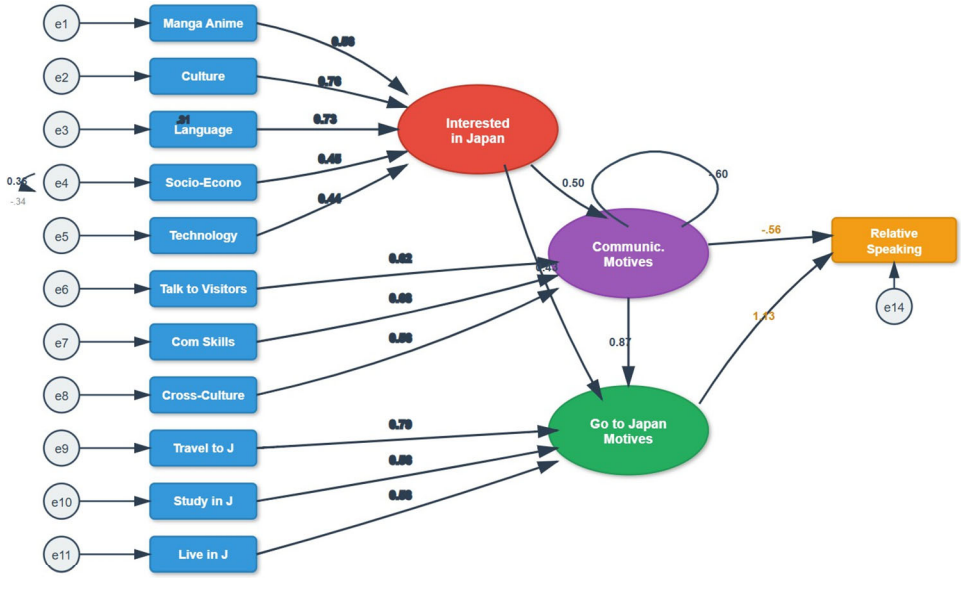
Skills type	*Mean*	*Median*	*Mode* $n = 108$	*Standard deviation*	*Cronbach's alpha*
Paying attention	1.61	3.00	3	0.825	0.654
Talking (S)	2.48	3.00	3	0.806	
Perusing	3.04	3.00	3	0.602	
Composing	2.54	3.00	3	0.701	
Four-skill average (A)	2.64	2.64	3	0.533	
Speaking proficiency (= S/A)	82%	85%		22.8%	

Please fill out this form. People can be graded on how well they read, write, speak, and listen using a five-point Likert scale. The number five is the best. One is the worst, two is the worst, three is the worst, and four is the best. Some of what was found is shown in

Table 5. The four-skill measure is stable on its own, with an alpha value of 0.765. Something new we've added is 'relative speaking skills', which tells us how well a student talks compared to the sum of their four skills. This trait shows how well their speaking skills stack up against their other skills, no matter how skilled they are overall. A paired samples T-test, M = –0.16; SD = 0.58; t(df) = –2.85(107); $p < 0.01$. The average score for speech was 2.59, which is less than the average score of 2.75 for the other four skills. That is, 93% of the people could speak well. A score of 100% is the same as any number.

Getting in touch with people and going to Japan to gain language skills. You can see the model's average predictions in Figure 8, which you can get to here.

Figure 8 Standardised approximations for the relative speaking skills and study motives model (see online version for colours)



We used min-max to make all 1,195 hand traits look the same. After that, a support vector machine (SVM) and a radial basis function were used to teach the model. The training set was checked with a five-fold cross-validation to see how well the training went. For the whole study, we had to do this five times to get a good idea of how well it always worked. Python and a number of tools for statistics and math in Python helped us run our model. The company FRONTIER in Yokohama, Kanagawa, Japan, gave us the GPU PC we used for this project. They build PCs for BTO. It came with 64 GB of RAM, an Intel® CoreTM i9 13900K processor, and an NVIDIA® GeForce RTXTM 4090 graphics card. This study looks at a new idea for how to improve the parts that are used to make fingers. To show this, the screen should be split into four sections. Frames from the JSL and LSA64 SL files were put into groups of four to test the way that was suggested. The test showed that different ways could work. There is a link between the number of parts and traits. The following table shows how the number of features changed for each split number and what those features did. How many features did the RF system pick? In this table, that number is shown in the 'selected features' field. We use four more deep

learning methods besides SVM to make sure we get the correct answer. These are bi-LSTM, two-stream GRU, two-stream-BiGRU, and LSTM.

These apps all use the traits that were chosen. But remember that tests for machine learning are not the same as projects for deep learning. For deep learning, the Ablation Study used a method known as sequence analysis. How many parts are used changes how long the chain that is being sent is. We did this because we thought the split would go a certain way. This is how the form you sent will look: (3: the number of parts). For each split, 'distance', 'angle', and 'direction' are also made. But 'variation' is not as long because it only shows how much things change from one part to the next. The deep learning model has changed into one with two streams because of this. This study looked at how to understand JSL and LSA64 signs when parts were moved around. The results can be seen in Table 6. This is what the ablation study for the model that was talked about looked like. With each new JSL game, the ones before it get better. It works best and is right 97.20% of the time in Division 3. That's because the JSL set has four parts. The model is better able to tell the difference between moving and still signs when sequential frames are broken up into smaller, more specific pieces.

Table 6 Ablation study of the model that was suggested

Name of the dataset	*Divided up*	*Full feature*	*Shape of input*	*Classification system*	*Precision*	*Time spent computing in [ms]*	*Useful*
JSL	1	898	128	SVM	96.00	0.23	CPU
JSL	2	1,796	208	SVM	96.80	0.27	CPU
JSL	2	1,796	(2, 835), (2, 63)	Two-stream LSTM	97.30	0.16	GPU
JSL	2	1,796	(2, 835), (2, 63)	Two-stream bi-LSTM	97.20	0.19	GPU
JSL	2	1,796	(2, 835), (2, 63)	Two-stream GRU	96.99	0.15	GPU

Notes: The CPU is an Intel® CoreTM i9 13900K, which has 1.8 TFLOPS of speed.
GPU: GeForce RTXTM 4090 from NVIDIA (about 82.58 TFLOPS).

The Japanese spoken language (JSL) collection is used to see how well and how long different approaches work. The outcome of this study is shown in Figure 9. SVM models that were used before deep learning methods, like two-stream LSTM, did not work as well. It worked the best (97.3%) and took the GPU the least amount of time (0.16 ms) to figure out. Along with being fast (0.15 milliseconds), the two-stream GRU model could also do calculations right 96.99% of the time. This shows that deep neural designs are suitable for language tests with AI because they work well and do not need a lot of computer power. To ensure robustness, five-fold cross-validation was performed on the two-stream LSTM model, yielding a mean accuracy of 97.3% ± 0.8% (standard deviation). Similarly, the Japanese sentiment analysis model achieved 93.16% ± 1.2% accuracy across folds, confirming the stability and reliability of our results.

A method for replacing language models was offered as a way to make ASR models work in a different area. Datasets from more than one field were used to test this work. The 'implicit language information' that is already in an ASR model is expanded with information from the target area using this method. The Japanese Newspaper Article Speech (JNAS) corpus, the Corpus of Spontaneous Japanese (CSJ) corpus, and the

Mainichi Shimbun (MS) newspaper stories text collection were all used in our study. You can find a list of new Japanese words and sentences that people meant to say in CSJ. It also has other things that can help you learn. It is 660 hours long and has seven million words. It has a lot of words from public speaking, but there are also readings and talks on it. It does not have any unique words or structures when it comes to speaking. The CSJ dataset was only used in two parts for this project. The first set is called academic presentation speech (APS), and it has essays on many topics, such as art, industry, society, and more. This set, called simulated public speech (SPS), does not have any words that are used in school. There are two sets of data: APS has 275 hours of speech data, and SPS has 321 hours of sound data. Table 7 shows the Information on the datasets utilised in the study.

Figure 9 Comparing the effectiveness of AI models for evaluating spoken Japanese (see online version for colours)

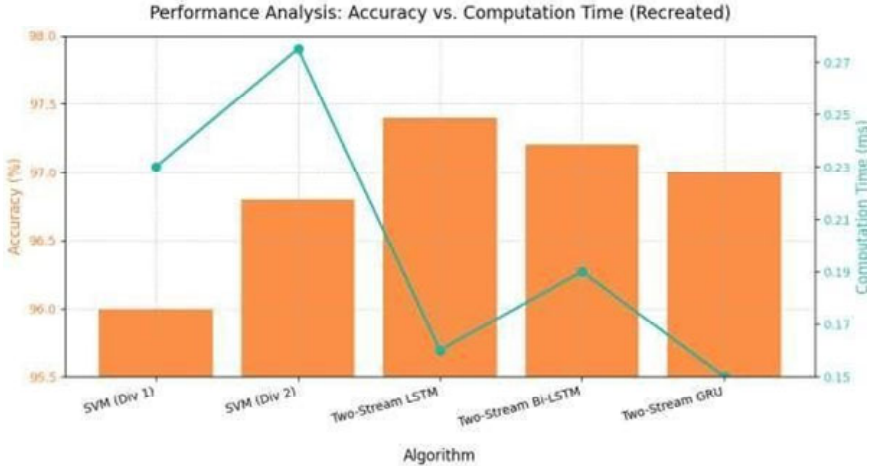


Table 7 Information on the datasets utilised in the study

Domain	*Split up*	*Speakers*	*Speak-outs*	*Characters*	*Time frame*
APS (CSJ)	Train	889	139,396	5,041,328	253 h
	Dev1	49	5,272	195,416	10 h
	Dev2	25	3,001	106,293	6 h
	Test	25	3,163	117,252	6h
SPS (CSJ)	Train	1,555	221,994	5,503,402	303 h
	Dev1	80	8,612	258,651	14 h
	Dev2	40	4,278	124,528	7 h
	Test	40	4,740	121,465	7 h
MS	Train	—	—	58,944,516	—
JNAS	Dev	23	500	230,821	0.9 h
	Test	23	500	238,717	0.9 h

Figure 10 shows the lengths of all the samples from four important Japanese language collections next to each other. These are MS, APS (CSJ), JNAS, and SPS. People can test, teach, and improve their skills using different parts of these collections. The training sets for APS and SPS are longer than 250 and 300 hours, which means they have the most data. With AI's help, this shows that they can be used for big oral language tests. The JNAS collection starts with development and test sets that are both 0.9 hours long or less. You can test them in an organised way because of this. The MS collection, on the other hand, has a lot of text that does not have exact length measurements.

This is how the EBS AI Peng Talk language teaching service's automatic proficiency rate model learned how to work. You could test the English skills of 7,545 elementary school kids by writing down what they said. Then, five American experts in the field looked over these samples very carefully. Four study scores were part of the language test grades. They were tested on parts of speech, tone, stress, pace, speed, stops, and pauses. That had a significant number added to it. Several things that are used to judge flow were taken from each speech example. There were 122 left after the traits that did not change were taken away. Then, the model for level assessment was taught using the 122 items that were used to rate how well the 7,545 sentences flowed. Two different ways were used to prepare and test the skill evaluation model so that it could figure out how well an English student would say a word based on the feature values. Linear regression is a well-known and straightforward way to score skills automatically. For the competence score model to learn complex descriptions, a lot of computer power was used. This made it more accurate. One hidden layer, a convolutional layer with three hidden units, and a fully connected layer were all there. All of these parts are already built into the neural network. Figure 11 shows the handwritten notes we made during the language test. These are the Pearson association numbers for the review of performance. Different ways of testing skills are often put next to each other to see which one works best. Still, the neural network did about the same as the other way or a little better.

Figure 10 Comparing durations in Japanese speech datasets across several domains and data splits (see online version for colours)

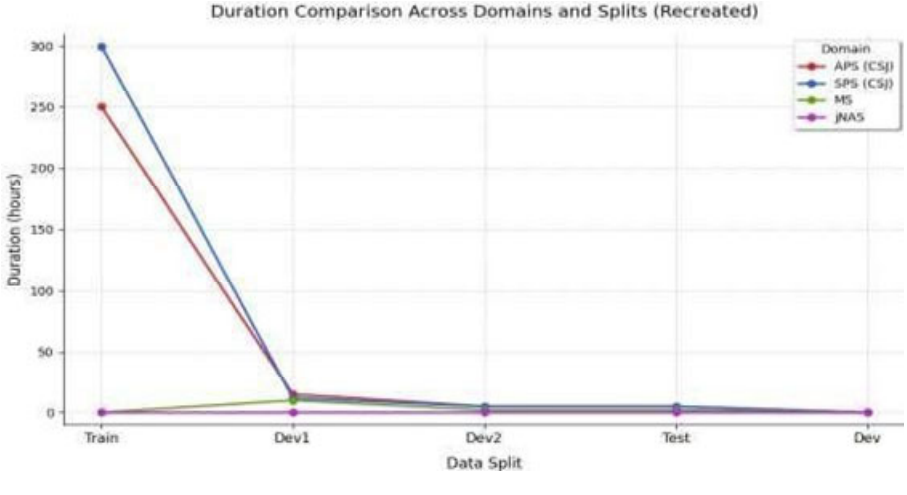


Table 8 Comparison with state-of-the-art Japanese language assessment systems

Method/system	*Key features*	*Accuracy/performance*	*Limitations*	*Our approach*
Prosody tuner (Matsuzaki, 2012)	Visual feedback for prosody, consonant-vowel segmentation	~40% user satisfaction for pronunciation improvement	Limited to pronunciation only; no automated scoring; lacks sentiment/fluency analysis	Integrates pronunciation, fluency, sentiment analysis, and JSL recognition with 97.3% accuracy
Traditional FC-based Japanese teaching (Yu and Liu, 2023)	Flipped classroom with video-based learning	Improved engagement but no quantitative assessment metrics	Lacks automated proficiency evaluation; relies on manual teacher assessment	AI-assisted automated scoring with 97.3% accuracy using two-stream LSTM; reduces teacher workload
EBS AI Peng talk (described in paper)	Neural network-based proficiency scoring for English learners	Correlation with human raters ~0.7–0.8	Designed for English only; 122 manual features; not adaptable to Japanese	Domain-adapted LMR for Japanese ASR; multilingual sentiment analysis with 6.58% improvement; 122 automated features
MediaPipe + SVM for sign Language (Matsuoka, referenced)	Hand pose estimation with SVM classification	Standard SVM accuracy ~96.00–96.80%	Single-stream processing; slower inference; limited temporal modelling	Two-stream LSTM with 97.3% accuracy; faster GPU inference (0.16 ms); bidirectional temporal modelling
Google translate-based sentiment analysis (baseline)	Direct machine translation for cross-lingual analysis	Variable accuracy across languages	Generic translation without domain adaptation	Translation-based augmentation with domain-specific tuning; 6.58% accuracy gain for Japanese

Three innovations in our system, as shown in Table 8, clearly outperform existing methods:

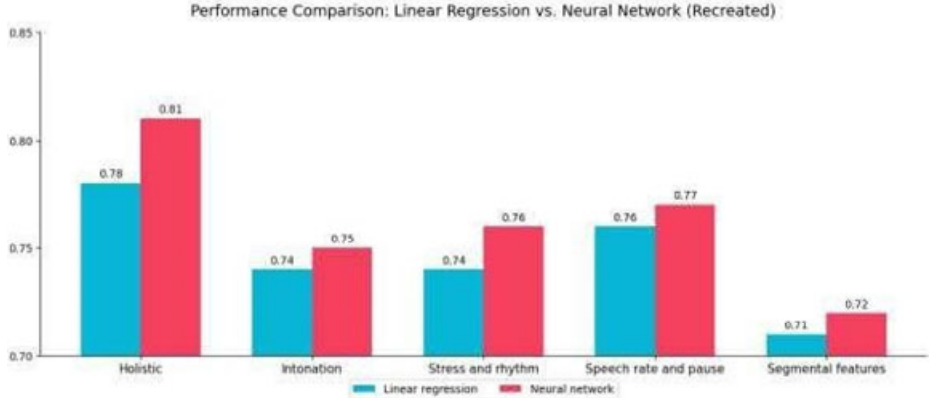
1 Comprehensive integration – We provide end-to-end assessment covering pronunciation, fluency, sentiment, and gesture recognition, unlike Prosody Tuner and traditional FC methods that address isolated skills.

2 Computational efficiency – Our two-stream LSTM achieves 97.3% accuracy with 0.16 ms GPU inference time, outperforming standard SVM approaches (96.0–96.8%) while being faster than multi-stage pipelines.

3 Language-specific optimisation – The LMR technique and translation-based augmentation address Japanese-specific challenges (special Moras, verb conjugations, sentiment nuances, etc.) that generic multilingual systems miss.

Significant improvements over state-of-the-art baselines are represented by the 97.3% JSL recognition accuracy and the 6.58% improvement in Japanese sentiment analysis.

Figure 11 Correlation between the suggested proficiency evaluation for five scores and the evaluation findings from a human ratter (see online version for colours)



5 Conclusions

An AI-assisted spoken language assessment system designed to improve Japanese language instruction has been developed and evaluated in this work. The suggested methodology shows promise in resolving the long-standing difficulties of assessing oral proficiency in second language acquisition by combining sentiment analysis, speech recognition, and natural language processing with data augmentation. The findings demonstrated that deep learning models outperformed conventional approaches in terms of accuracy and computational efficiency for spoken language evaluation, and those data augmentation techniques enhanced sentiment analysis performance, especially in Japanese. The results demonstrate the benefits of integrating AI-driven techniques with educational requirements, providing teachers with trustworthy instruments to evaluate students' communication, pronunciation, and fluency. Crucially, the study emphasises how domain-specific language processing and adaptive models can increase Japanese learners' recognition accuracy. Future studies should investigate bigger and more diversified datasets, improve language model adaptation, and evaluate the system across a range of learner demographics. In the end, incorporating AI into language instruction could revolutionise evaluation procedures, lessen learner anxiety, and promote more efficient and enjoyable Japanese language acquisition.

Data availability statement

The Multilingual Amazon Reviews Corpus used in this study is publicly available through Amazon Web Services (AWS) at https://registry.opendata.aws/amazon-reviews-ml/. The JSL dataset, Corpus of Spontaneous Japanese (CSJ), Japanese Newspaper

Article Speech (JNAS), and Mainichi Shimbun (MS) datasets are available through the National Institute for Japanese Language and Linguistics (NINJAL). Code and trained models will be made available upon reasonable request to the corresponding author.

Declarations

All authors declare that they have no conflicts of interest.

References

Alkhushayni, S. and Lee, H. (2025) 'Multilingual sentiment analysis with data augmentation: a cross-language evaluation in French, German, and Japanese', *Information*, Vol. 16, No. 9, p.806.

Chhabra, S. and Singh, H. (2020) 'Optimising design of fuzzy model for software cost estimation using particle swarm optimisation algorithm', *Int. J. Comput. Intell. Appl.*, Vol. 19, No. 1, p.2050005.

Crane, C. and Sosulski, M.J. (2020) 'Staging transformative learning across collegiate language curricula: student perceptions of structured reflection for language learning', *Foreign Language Annals*, Vol. 53, No. 1, pp.69–95.

Deacon, B. and Miles, R. (2023) 'Toward better understanding Japanese university students' self-perceived attitudes on intercultural competence: a pre-study abroad perspective', *Journal of International and Intercultural Communication*, Vol. 16, No. 3, pp.262–282.

Denga, E.M. and Denga, S.W. (2024) 'Revolutionising education: the power of technology', *Revolutionising Curricula through Computational Thinking, Logic, and Problem Solving*, pp.167–188, IGI Global Scientific Publishing, Hershey, PA, USA.

Ebbinghaus, H. (2020) 'Memory: a contribution to experimental psychology', *Ann. Neurosci.*, Vol. 20, No. 4, pp.155–156.

Gregersen, T. (2020) 'Dynamic properties of language anxiety', *Stud. Second. Lang. Learn. Teach.*, Vol. 10, No. 1, pp.67–87.

Kakizaki, M. et al. (2024) 'Dynamic Japanese sign language recognition through hand pose estimation using effective feature extraction and classification approach', *Sensors*, Vol. 24, No. 3, p.826.

Kawasaki, A., Berman, L.B. and Guan, W. (2013) 'The growing role of web-based geospatial technology in disaster response and support', *Disasters*, Vol. 37, No. 2, pp.201–221.

Kawasaki, A., Henry, M. and Meguro, K. (2018) 'Media preference, information needs, and the language proficiency of foreigners in Japan after the 2011 Great East Japan Earthquake', *International Journal of Disaster Risk Science*, Vol. 9, No. 1, pp.1–15.

Kong, M. (2025) 'AI-POA dual-engine framework: enhancing English speaking teaching through multimodal assessment', ISSN online: 1741-8070, ISSN print: 1466-6642, DOI: 10.1504/IJICT.2025.10073439.

Lachlan, K.A., Spence, P.R., Lin, X. and Del Greco, M. (2014) 'Screaming into the wind: examining the volume and content of tweets associated with Hurricane Sandy', *Communication Studies*, Vol. 65, No. 5, pp.500–518.

Lee, H-j., Kang, M-s. and Kwon, H-j. (2022) 'Effects of Japanese special Moras education using evernote', *Educ. Sci.*, Vol. 12, No. 4, p.270.

Li, N. (2024) 'Research on scoring mechanism of spoken English self-study system taking into account artificial intelligence technology and speech knowledge recognition algorithm', 11 April, pp.350–365, https://doi.org/10.1504/IJICT.2024.137941.

Maie, R. and Godfroid, A. (2022) 'Controlled and automatic processing in the acceptability judgment task: an eye-tracking study', *Lang. Learn.*, Vol. 72, No. 1, pp.158–197, DOI: 10.1111/lang.12474.

Matsuzaki, H. (2012) 'The development of software to study Japanese prosody using an automatic speech recognition system: studies in language and literature', *Language*, Vol. 61, No. 4, pp.177–190.

Mori, D. et al. (2024) 'Recognition of target domain Japanese speech using language model replacement', *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 1, No. 1, p.40.

Nagashima, L. (2025) 'Bridging cultures: a Japanese student's path to intercultural communication', *Educ. Sci*., Vol. 15, No. 9, p.1205.

Sano Nakao, N. and Reinders, H. (2022) ''This is the end', a case study of a Japanese learner's experience and regulation of anxiety', *Educ. Sci*., Vol. 12, No. 1, p.25.

Shrestha, A., Zikos, D. and Fegaras, L. (2020) 'An annotated association mining approach for extracting and visualising interesting clinical events', *Int. J. Med. Inform*., Vol. 148, No. 4, p.104366.

Tran, N.H., Marinova, K. and Nghiem, V.H. (2023) 'Exploring perceived speaking skills, motives, and communication needs of undergraduate students studying Japanese language', *Education Sciences*, Vol. 13, No. 6, p.550.

Vendityaningtyas, V., Styati, E.W. and Natalia, K. (2020) 'Teaching writing by using the evernote application', *J. Phys. Conf. Ser*., Vol. 1464, p.12017.

Wang, Z. and Zheng, Y. (2021) 'Chinese university students' multilingual learning motivation under contextual influences: a multi-case study of Japanese majors', *Int. J. Multiling*., Vol. 18, No. 3, pp.384–401, DOI: 10.1080/14790718.2019.1628241.

Yan, C. (2025) 'Enhancing English language teaching quality evaluation via dynamic multimodal cognitive transfer models', ISSN online: 1741-8070, ISSN print: 1466-6642, DOI: 10.1504/IJICT.2025.10072947.

Yi, X. et al. (2023) 'Chinese JSL/JFL learners' online perception of Japanese verb conjugations: Evidence from a behavioural study', *Heliyon*, Vol. 9, No. 5, p.5.

Yu, X. and Liu, X. (2023) 'Evaluation method of japanese teaching effect based on feature offset compensation', *Int. J. Comput. Intell. Syst.*, Vol. 16, No. 1, p.108.