Semi-development and Performance Evaluation of a Pronunciation Judgment System Using Free Machine Learning Services

Yukinobu Miyamoto

Department of Information Engineering, Faculty of Science and Engineering,
Otemon Gakuin University, Osaka, Japan

*Email: miyamoyu@gmail.com

Abstract

With the recent development of the global community, communication skills in languages other than one's native tongue have become essential. While attending school is an effective method of language learning, it is often difficult due to time and financial constraints. If it becomes possible to acquire language pronunciation through self-study, it is expected that the learning process can be significantly shortened. Furthermore, building such a practice system without specialized IT knowledge, such as programming, would be a great benefit to educational settings. This paper describes a pronunciation assessment system for language learning that utilizes a free machine learning service. The target language is Japanese. By providing machine learning with speech data of homonyms that are difficult for non-native speakers to distinguish, we build a system that can assess the accuracy of pronounced words. For machine learning, we use Google Teachable Machine, a free service that allows system building without specialized IT knowledge. Experiments using this method demonstrate that we have constructed a system that can assess the accuracy of native speaker pronunciation with a very high probability.

Keywords

Language Learning, Japanese Language, Pronunciation, Classification, Google Teachable Machine

Introduction

While grammar and vocabulary instruction are central to language acquisition, pronunciation training tends to be neglected. However, accurate pronunciation forms the foundation for smooth communication and is crucial for avoiding misunderstandings and sustaining dialogue. Particularly in second language learning, inaccurate pronunciation makes it difficult for learners to convey their intended meaning correctly, often leading to stagnation in conversation. Therefore, pronunciation acquisition is positioned as an indispensable element for improving overall language proficiency.

Submission: 29 November 2025; **Acceptance:** 29 December 2025; **Available online:** December 2025

The significance of pronunciation is also clear from the perspective of communicative clarity. Even if a learner constructs grammatically correct sentences, inaccurate pronunciation makes it difficult for listeners to interpret the meaning correctly. Furthermore, since the phonological systems of languages differ significantly, learners are susceptible to the influence of their native language and must consciously strive to acquire accurate pronunciation. Moreover, pronunciation learning contributes to improving listening skills. From the perspective of "reconstructing auditory perception," where sounds learners can produce are also easier to recognize aurally, pronunciation practice is suggested to aid not only speaking but also the development of listening skills. Furthermore, gaining confidence in pronunciation is important as it increases learners' motivation to speak and promotes active communication behavior.

Against this backdrop, pronunciation teaching methods utilizing machine learning have gained attention in recent years. Traditional teaching methods relied on individual feedback from instructors, making it difficult to ensure sufficient practice opportunities. However, advances in speech recognition technology have enabled learners to receive immediate feedback on pronunciation differences. Technologies like Google Speech-to-Text and Apple Siri analyze learner speech and provide error feedback. Furthermore, deep learning-based pronunciation evaluation systems quantify subtle errors by comparing speech to native audio, providing precise feedback. Additionally, AI-powered pronunciation correction applications, such as ELSA Speak and SpeechAce, are becoming increasingly prevalent, and the development of training systems incorporating unsupervised learning and reinforcement learning is advancing.

Numerous computer systems designed for pronunciation assessment and practice have been developed to date. Much of this research has focused on support for the hearing impaired and pronunciation education for international students. For speakers struggling with accurate pronunciation in their native language and for non-native speakers learning a foreign language, pronunciation plays a crucial role alongside grammar and vocabulary in enabling mutual understanding.

This research aims to build a pronunciation support system that language educators can easily utilize without requiring specialized knowledge or high costs, by applying such machine learning technologies. Specifically, it leverages free machine learning services to evaluate their performance and verify their effectiveness for foreign language learners. These services are available at no cost and can be operated entirely online, making them accessible even to educators without programming knowledge. Furthermore, since the learning models can be converted into Python or JavaScript formats as needed, collaboration with IT specialists enables the development of more advanced systems.

Conventional commercial systems relied on fixed pronunciation models, resulting in limited flexibility. In contrast, the approach adopted in this research allows the machine learning models to be trained using data collected by users. This provides scalability, enabling support for subjects not covered by existing models, such as dialects and minority languages. The tasks required of users are limited to data collection and organization, eliminating the need for conventional development processes. This approach is termed "semi-development" in this research.

Therefore, this research aims to build a pronunciation education support system combining three key features: scalability, flexibility, and low cost. By integrating pronunciation education with artificial intelligence technology, it presents a new learning environment and proposes an effective and sustainable approach to language learning.

Methodology

- **Outlines of the proposed system**

This section outlines a pronunciation assessment system for language learning support utilizing machine learning. The general procedure for machine learning involves constructing a learning model using reference training data, then validating it with test data to evaluate its performance. Subsequently, the model, once validated, is used to assess the accuracy of unknown data not included in the training. This research applies this procedure to language learning. We aim to construct a system capable of evaluating the accuracy of judgments for unknown speech by first recording reference word audio and then building a learning model.

This study adopts Japanese as the model language and aims to establish a methodology for constructing and operating a system that determines whether pronunciations outside the training data fall within the pronunciation range of the same word, without requiring specialized knowledge or advanced technology. The Japanese words targeted are those with identical hiragana spellings (Japanese language letter) but differing meanings due to variations in accent placement; these are referred to as “homophones” in Japanese. While homophones may change accent depending on context, this study focuses on the standard accent when the word is taken independently.

Based on the above definition, Japanese words for model construction are collected from native Japanese speakers and trained using machine learning. After training completion, the system's validity is first evaluated using the training data. If validity is confirmed during this process, additional verification is performed on untrained test data to measure its judgment accuracy. The test data consists of audio recordings from native Japanese speakers not included in the training data, evaluating whether the given pronunciation falls within the scope of the same word. Although Japanese was used as the training language, the system is designed to function similarly for other languages by replacing the language and speaker nationality.

- **Free machine learning services used in this research**

This section provides an overview of Google Teachable Machine (GTM), a machine learning classification tool. GTM is a web-based tool powered by Google's deep learning infrastructure. Its key feature is the ability to easily create machine learning models using data such as images, audio, and poses, even without programming knowledge. Figure 1 shows part of GTM's startup screen. The figure illustrates the learning process, including data collection, model training, and model export.

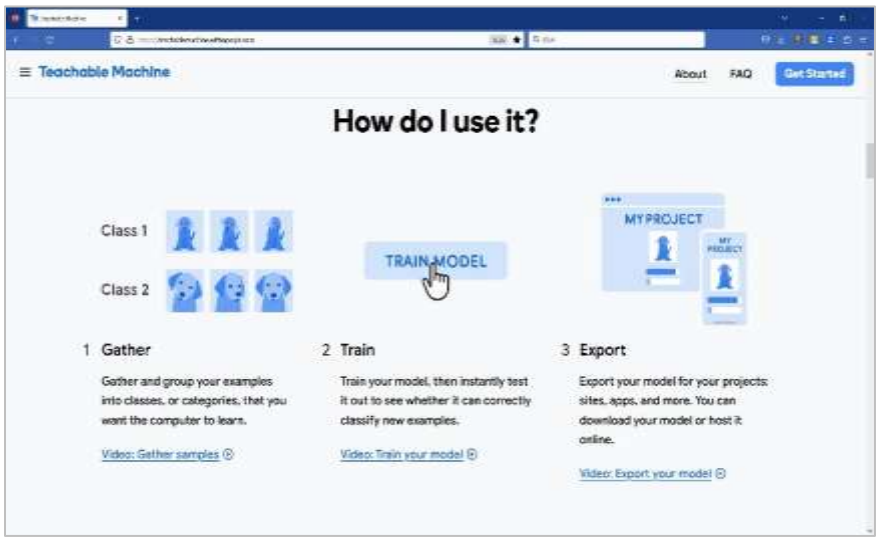

Figure 1. How to use Google Teachable Machine

Figure 1 shows how to use Google Teachable Machine. First, select the type of data to be learned. The available data types are images, audio, and poses. Next, collect the selected data. For images, capture via camera or acquisition from the internet is possible. For audio, recording via microphone input or collection from the internet, similar to images, is available. For poses, video capture via camera is considered the standard method. Once data collection is complete, you can configure classification and training settings within GTM to build a learning model. The constructed model can be executed immediately in a web browser and can also be downloaded for integration into external applications or websites. Additionally, a framework of control codes is partially provided in Python and JavaScript formats for use at this stage.

After the learning model is complete, test data can be used to verify which of the learned classes the data belongs to. For images, dragging and dropping a test image onto the browser displays the probability of its class membership as a numerical value and a bar graph. For audio, the current implementation does not provide a file input function; verification is only possible via real-time input through a microphone. When using pre-recorded audio, playback via a media player and subsequent capture through a microphone enables verification. Direct speech input is also possible, but reproducing identical audio is difficult, making it disadvantageous for ensuring reproducibility. Therefore, this research employs a method where pre-recorded audio is played back through a speaker and then re-input via a microphone, enabling multiple verifications using the same audio.

GTM is one of the useful tools for deepening the fundamental understanding of machine learning. It enables the easy construction of machine learning models and their deployment for diverse applications without requiring specialized knowledge in IT or programming. Furthermore, as it is provided as a web service, once the data for classification is prepared, all steps can be completed solely through mouse operations in a browser, offering excellent usability. Moreover, for scenarios requiring advanced analysis or integration into applications, the model can be converted into a programming language format, enabling flexible utilization tailored to the user's proficiency level. In this research, we apply GTM to speech recognition and attempt to build a pronunciation judgment system.

Results and Discussion

- **Outlines of experiments**

First, we describe the experimental procedure used in this study. For the experimental environment to record audio data, we prepared a PC connected to a microphone in a quiet room. In GTM's audio project, it is essential to record background noise for at least 20 seconds as the initial class for classification. Therefore, we recorded 20 seconds of noise data in a quiet state. After completing the noise recording, GTM automatically divides the recorded data into 1-second segments and saves them as 20 audio data files.

Next, audio recordings of the Japanese words to be classified were sequentially captured. Recording began with audio from Japanese native speakers, which would be used as training data. The words targeted in this experiment are the 10 types shown in Table 1, which form pairs of two words each, listed sequentially from top to bottom. All these pairs are homophones. For example, “hashi” shown in the top row of Table 1 means “bridge” when the accent is on the first syllable and “chopsticks” when the accent is on the second syllable. In Table 1, an accent mark is added to the relevant syllable to emphasize the accent position. The 10 types of words were recorded at 1-second intervals over 10 seconds per subject, and this process was repeated for all subjects.

Table 1. Japanese words and meanings for the experiments

Japanese Word	Meaning	Accent
hashi	bridge	[$\overline{ha}\ \underline{shi}$]
	chopstick	[$\underline{ha}\ \overline{shi}$]
kumo	cloud	[$\overline{ku}\ \underline{mo}$]
	spider	[$\underline{ku}\ \overline{mo}$]
kaki	oyster	[$\overline{ka}\ \underline{ki}$]
	persimmon	[$\overline{ka}\ \overline{ki}$]
momo	thigh	[$\overline{mo}\ \underline{mo}$]
	peach	[$\overline{mo}\ \overline{mo}$]
ika	less than	[$\overline{i}\ \underline{ka}$]
	squid	[$\overline{i}\ \overline{ka}$]

After recording the speech data of all subjects, we constructed the training dataset. Specifically, we first extracted the audio for “hashi (bridge)” and created a new class following background noise, registering all participants' audio data. We applied the same procedure to the other nine words, creating a total of ten classes of training data. Once the training data classes were organized, we executed training in GTM to generate the learning model. The training process was completed in a few minutes at most.

After training completion, testing was conducted using the trained model. First, for performance evaluation, audio from Japanese native speakers (the training data) was played back, and the probabilities assigned by GTM to each homophone were recorded. Since audio is time-series data and output values fluctuate depending on playback timing, we adopted a best-effort evaluation method, treating the best value from multiple trials as the result for that data point. This

procedure was repeated for all training data, and the average value based on the number of subjects was used as the classification accuracy for each word.

After completing testing with the training data, untrained audio data was played sequentially, and testing was performed using the same method. Figure 2 shows an example screen from the test phase. This figure shows the result of the classification experiment for “hashi (bridge)” and “hashi (chopsticks)”, where when the word “hashi (bridge)” was spoken, GTM judged it as “bridge: 100%, chopsticks: 0%”. For the test data, similar to the training data, the average value based on the number of subjects was used as the final classification accuracy for each word.

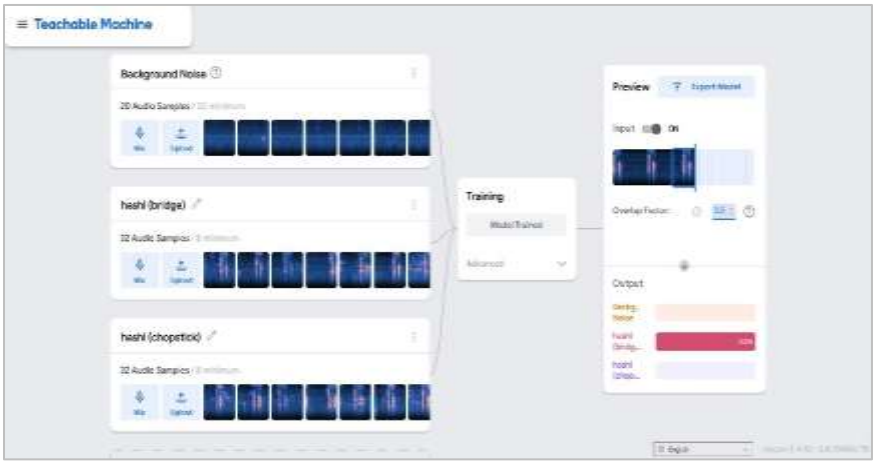

Figure 2. Example of the experimental results in test phase

Furthermore, the attributes of the experimental data are shown in Table 2. Table 2 indicates the recording period, total number of subjects, gender breakdown, and average age. The audio data was recorded from May to July 2025, with a total of 32 participants (25 males and 7 females), and the average age of the subjects was 18.97 years.

Table 2. Attribute of the samples for the experiments

Nation	Month/Year	Number of Test Subjects	Age
Japanese	05-07/2025	32 (Male:25 / Female:7)	18.97

- **Experimental results**

Table 3 shows the numerical results of the experiment. First, the classification results based on the training data - speech from Japanese subjects - achieved an average classification accuracy of 94.6% across all words. Examining the accuracy for each word individually reveals that even the lowest value was 89.8%, indicating that accuracy above 90% was generally maintained.

Next, regarding the classification results based on the audio used as test data, an average classification accuracy of 87.5% was achieved across all words. Similarly, when verifying accuracy on a word-by-word basis, the lowest value still reached 83.3%, indicating that overall classification accuracy above 83% was maintained. Furthermore, focusing on variance, it was confirmed that the variance in the test data did not significantly diverge from that of the training data.

Table 3. Experimental results

Japanese Word	Meaning	Training Samples		Test Samples	
		Ave.	*SD*	*Ave.*	*SD*
hashi	bridge	89.8%	0.1143	85.5%	0.0754
	chopstick	97.6%	0.0451	91.2%	0.0680
kumo	cloud	96.2%	0.0517	83.3%	0.1066
	spider	92.5%	0.0685	88.6%	0.0656
kaki	oyster	96.3%	0.0439	88.4%	0.0777
	persimmon	95.8%	0.0422	93.1%	0.0602
momo	thigh	93.2%	0.0513	88.0%	0.0923
	peach	92.6%	0.0814	89.4%	0.0926
ika	less than	96.5%	0.0456	92.5%	0.0579
	squid	95.2%	0.0444	88.5%	0.0822
Classification Accuracy		94.6%	0.0588	88.9%	0.0778

Discussions

This section presents an analysis based on the experimental results. From the results in Table 3, experiments using training data achieved an overall classification accuracy of approximately 95%, which can be evaluated as a sufficiently good result. Furthermore, when examined word by word, the minimum value was 89.8%, confirming that accuracy above 90% was generally maintained. Generally, in machine learning, it is desirable for accuracy on training data to approach 100%. However, since speech data is dynamic, time-series-based data, factors such as speech playback timing may cause it to be misclassified as non-training data. Therefore, compared to static data like images, it is anticipated that classification accuracy may not always reach high values. Considering this point, the 95% accuracy achieved in this experiment is considered a sufficiently valid result for utilizing GTM as a speech classification system.

Next, in experiments using test data, a classification accuracy of 88.9% was achieved, which can also be evaluated as a favorable result. While this value is slightly lower compared to the training data, it is a natural phenomenon in machine learning for test data accuracy to decrease relatively. Within this context, obtaining value close to the training data can be positively evaluated. When examined word by word, “kumo (cloud)” had the lowest value at 83.3%, but overall, a classification accuracy of 83% or higher was maintained. Furthermore, no significant divergence in variance was observed between the training and test data. Based on these results, it is considered that the pronunciation capable of being appropriately judged as Japanese was sufficiently realized overall in this experiment.

Note that the Japanese audio samples used in this experiment were recorded in a specific region within Japan (Osaka) and thus may contain tendencies slightly different from standard Japanese pronunciation. Therefore, the verification results obtained in this experiment should be positioned as evaluating “whether the pronunciation is faithful to the Japanese audio samples used as reference” rather than “whether correct Japanese pronunciation is being produced.”

Conclusion

This paper reports on the development and performance evaluation of a prototype simple pronunciation assessment system for language learning. In constructing the system, we leveraged existing classification services based on machine learning and implemented design features to ensure ease of use even for those without expertise in information science. Experiments validated the system's effectiveness using Japanese native speaker audio as training data and distinct native speaker audio as test data. The results showed an accuracy of 94.6% on the training data, confirming that the system possesses sufficient judgment capability. Furthermore, it achieved an accuracy of 88.9% on the test data, demonstrating that it functions effectively even on unknown data.

Future challenges include expanding the variety and scale of words used in the training data. It must be acknowledged that the collection of training data in this study was limited to five types of Japanese homophones, preventing verification across a broad range of Japanese pronunciations. Furthermore, this verification targeted only native Japanese speakers, making it insufficient for evaluating the pronunciation of Japanese learners. Therefore, future work is expected to expand the scope of application to include pronunciation assessment for foreign language learners by broadening the variety of Japanese words and collecting audio data from speakers of other nationalities.

Acknowledgements

There is no grant or funding bodies to be acknowledged for preparing this paper. The authors acknowledge with gratitude the contributions of the researchers and authors whose published works were reviewed and analyzed in this study.

References

Apple Siri, Apple Inc. (2011). https://www.apple.com/siri/

ELSA Speak, ELSA Corp. (2016). https://elsaspeak.com/

Google Speech-to-Text, Google LLC. (2016). https://cloud.google.com/speech-to-text

Google Teachable Machine, Google LLC. (2017). https://teachablemachine.withgoogle.com/

Ito, M. (2021). Prospects for approaches to speech sound in Japanese language education: from pronunciation correction, speech instruction, and speech learning support to practice of voice-themed dialogue. *Waseda Studies in Japanese Language Education*, 30, 129-148. DOI: 10.15055/00007802

Katsuse, I. M. (2017). Support system for pronunciation teaching and practice in special education classes for language-disabled children enrolled in regular schools. *Transactions Japanese Society for Information and Systems in Education*, 34(1), 7-19. DOI: 10.14926/jsise.34.7

Nakamura, T. (1999). The “Ai-chan no Te” speech training system for aurally impaired children. *Technical Spoken Language Processing, The Special Interest Group Technical Reports of IPSJ*, vol. 25, no. 12, pp. 57-58. CiNii: 110002771589

Okazaki, Y., Matsunaga, S., Tanaka, H., & Watanabe, K. (2014). Development of a Japanese pronunciation practice support system adapted for foreign students’ Japanese levels and native

languages. *Japanese Society for Information and Systems in Education Research Report*, 28(7), 179-186. https://cir.nii.ac.jp/crid/1520009409478048128

Sasaki, K., & Miwa, J. (2015). An interactive system for pronunciation assessment of Japanese geminates using Android devices and its evaluation. *IEICE Technical Report*, ET2014-79(2015-1), 39-44. IEICE ID: 110009911910

SpeechAce, Speechace LLC. (2018). https://www.speechace.com/

Toda, T., Kinoshita, N. & Chris, S. (2006). Studies in the phonological acquisition process of second languages. *Report on the Research Achievements of Grants-in-Aid for Scientific Research*. KAKEN: 15320083

Tsubota, Y., Kawahara, T., & Dantsuji, M. (2001). English pronunciation instruction system using pair-wise discrimination between error patterns of Japanese speakers. *Proceedings of Meetings on Acoustics, the Acoustical Society of Japan*, 2001(2)2, 341-342. https://ci.nii.ac.jp/search?q=110003110344

Umezaki, T., Kuratani k., & Fujiyoshi, H. (1997). Speech training support system for hearing impaired children using the network environment. *The Transactions of the Institute of Electronics, Information and Communication Engineers*, J80-D-II (4), 925-932. 10.14923/transinfj.J80-D-II.925