Comparing Automatic Speech Recognition and Teacher Assessments of Japanese Learners' Mandarin Chinese Pronunciation: Accuracy, Agreement, and Pronunciation Difficulty Detection
(自动语音识别与教师对日本汉语学习者普通话发音评估的比较：准确性、一致性及发音困难识别)

Huang, Weihsun
(黄暐勋)
Kobe University
(神户大学)
223c302@stu.kobe-u.ac.jp

Kashiwagi, Harumi
(柏木治美)
Kobe University
(神户大学)
kasiwagi@kobe-u.ac.jp

Kang, Min
(康敏)
Kobe University
(神户大学)
kang@kobe-u.ac.jp

Abstract: Computer-assisted pronunciation training (CAPT) increasingly incorporates automatic speech recognition (ASR) to provide pronunciation assessment and feedback. However, the extent to which ASR systems evaluate non-native Mandarin Chinese speech in a manner comparable to human teachers remains unclear. This study compares the assessments generated by three ASR systems—Whisper, Azure, and Gladia—with ratings provided by native Chinese-speaking teachers for the word-level Mandarin Chinese pronunciation of 31 Japanese learners. Two research questions are addressed: (1) To what extent do these ASR systems assess learner pronunciation comparably to teachers? (2) Can ASR assessments help identify learners' pronunciation difficulties? A three-point scoring scheme was developed to evaluate learners' productions of 20 Mandarin Chinese words. Comparative analyses were conducted from the perspectives of learner proficiency and pronunciation characteristics. The results showed that all three ASR systems generally underestimated learner performance relative to teacher ratings, although Whisper produced assessments that were most consistent with those of the teachers. The agreement between ASR and teacher assessments also varied according to learner proficiency. Furthermore, ASR performance was strongly influenced by initial–final combinations, suggesting that ASR assessments can help identify specific pronunciation difficulties. These findings support the potential of ASR as a complementary tool for pronunciation assessment in Mandarin Chinese CAPT.

摘要：随着计算机辅助发音训练（Computer-Assisted Pronunciation Training, CAPT）的发展，自动语音识别（Automatic Speech Recognition, ASR）系统日益广泛地应用于发音评估与反馈。然而，ASR 系统对非母语者普通话发音的评估能否达到与教师相近的水平，仍缺乏充分的实证研究。本研究比较 Whisper、Azure 和 Gladia 三种

ASR 系统与中文母语教师对 31 名日本学习者 20 个汉语词语发音的评估结果，以探讨：（1）ASR 系统在词语层面的发音评估与教师评分具有多大程度的一致性？（2）ASR 评估是否有助于识别学习者的发音困难？本研究建立三级评分标准，对学习者发音分别进行 ASR 评分与教师评分，并从学习者水平和发音特征两个层面比较两者的评估结果。研究结果显示，三种 ASR 系统均倾向于低估学习者的发音表现，其中 Whisper 与教师评分的一致性最高。此外，ASR 与教师评分的一致程度会因学习者水平而有所不同。进一步分析发现，ASR 评估结果受到声母—韵母组合的显著影响，表明 ASR 评估有助于识别学习者的具体发音困难。本研究结果支持 ASR 作为普通话计算机辅助发音训练中辅助发音评估工具的应用潜力。

Keywords: Automatic speech recognition, ASR Assessment, Japanese learner of Chinese, Teacher assessment, Computer-assisted pronunciation training (CAPT)

关键词: 自动语音识别，ASR 评估，日本汉语学习者，教师评估，计算机辅助发音训练（CAPT）

1.Introduction

Mandarin Chinese courses are offered as required electives at many universities in Japan. First-year students who begin studying Chinese as a foreign language are typically required to take two classes per week and continue their studies for 15 weeks during both the spring and fall terms. Owing to large class sizes, students often have limited opportunities for pronunciation practice, a challenge common to foreign language classroom instruction (Gao, 2025; Chen, 2011).

Computer-assisted pronunciation training (CAPT) has been recognized as an effective pedagogical approach for improving learners' pronunciation, particularly in the context of English as a foreign language (EFL) (Fouz-González, 2015). CAPT systems frequently incorporate automatic speech recognition (ASR) technology, which is used to detect phonetic errors, provide corrective feedback, ultimately enhance learners' pronunciation and potentially improve awareness of grammatical features (Ehsani & Knodt, 1998; Burleson, 2007; Neri et al., 2008; Eskenazi, 2009; Wang & Young, 2014; Tsai, 2019; McCrocklin, 2019; Dai & Wu, 2023; Issa & Hahn-Powell, 2025). ASR-based CAPT has also been applied to the learning of Chinese as a foreign language (CFL) (Da, 2015; Zhao et al., 2019; Watanabe et al., 2019, Li et al., 2024).

Despite these developments, the accuracy of ASR in evaluating non-native speech remains a matter of ongoing concern (Ehsani & Knodt, 1998, Derwing et al, 2000; Sunaoka, 2018; McCrocklin et al., 2019; Inceoglu et al., 2023; Hirai & Kovalyova, 2024). Eskenazi (2009) emphasized that CAPT systems employing ASR should be capable of detecting

individual pronunciation errors and assessing fluency in a manner comparable to human experts. Similarly, O'Brien et al. (2018) highlighted the need to identify ASR-derived measure that align closely with human judgements. Even with recent technological advances, the question of how accurately ASR systems recognize learner speech remains critical. A recent study investigating the performance of five speech-to-text applications for EFL learners has shown that recognition accuracy is influenced not only by the systems' technical capabilities but also by characteristics of the learners' utterances (Hirai & Kovalyova, 2024).

In the domain of CFL, Sunaoka (2018) analyzed the recognition accuracy of non-native speech in a Chinese long-distance group discussion using the ASR function integrated in Google Translation, arguing that teachers must verify ASR evaluations to compensate for technological limitations. Nevertheless, it remains unclear how accurately contemporary ASR technologies recognize CFL learners' speech, and which ASR-derived assessments most closely reflect human judgements. To address these gaps, the present study compares assessments of Japanese learners' speech generated by three ASR systems with evaluations provided by native Chinese-speaking teachers.

2. Literature Review

2.1 Computer-Assisted Pronunciation Training Systems with ASR

A substantial number of CAPT systems incorporating ASR have been developed for various EFL learning purposes. Burleson (2007) employed ASR to improve segmental errors produced by non-native speakers. Five Mandarin-speaking learners of English underwent pronunciation training targeting six phonemic contrasts. The ASR was used to recognize their productions and provide feedback, while native English listeners evaluated pre- and post-training recordings using forced-choice minimal pair tasks. The results demonstrated a significant improvement in learners' segmental intelligibility. Wang and Young's (2014) ASR-based iCASL system further examined the presentation of corrective feedback and demonstrated the effectiveness of a three-level feedback scheme. Windows Speech Recognition (WSR), a built-in speech recognition tool in Microsoft Windows, has also been integrated into sentence dictation practice, with findings suggesting that it can serve as a useful complement to face-to-face pronunciation instruction (McCrocklin, 2019). More recently, Issa and Hahn-Powell (2025) reported the use of a fine-tuned speech model within a CAPT system to investigate the effectiveness of ASR corrective feedback on the pronunciation of Arabic. These studies illustrate the diversity of ASR technologies employed in CAPT systems.

In the field of CFL, Da (2015) introduced Google's ASR, embedded in Chrome browser, into classroom pronunciation practice for ten non-native learners, suggesting that meaningful or frequently used expressions may be more suitable for Pinyin activities than isolated syllables groups. Zhao et al. (2019) developed "KoToToMo", a smart phone-based system for read-aloud practices that utilizes operating-system speech recognition. In addition to repetition and shadowing tasks, the system allows learners to conduct pronunciation "trials", enabling them to confirm their performance based on recognition

results and receive feedback. Watanabe et al. (2019) proposed the "ST-lab" system, which integrates both ASR and text-to-speech (TTS) technologies via the Web Speech API. In its "Reading Aloud Practice" module, the ASR component evaluates learners' pronunciations and displays the recognition results, allowing learners to repeat items until a "correct answer" is achieved or optionally skip them. As with EFL, a wide range of ASR technologies has been integrated into CAPT systems for CFL (Da, 2015; Wei & Zhang, 2018; Watanabe et al., 2019; Zhao et al., 2019).

2.2 ASR Assessment Accuracy of Learner Speech

The accuracy of ASR remains a critical issue in its deployment within CALL environments, largely because most commercial speech recognizers are trained on standard native pronunciations (Ehsani & Knodt, 1998). Derwing et al. (2000) evaluated the effectiveness of a widely used ASR package in providing corrective feedback based on two criteria: whether the ASR system recognizes ESL speech at an acceptable level and whether the ASR system has the potential to identify production difficulties. They argued that the usefulness of ASR depends on how closely its assessments of ESL speech approximate those of native listeners, and they emphasized the need for careful evaluation of ASR applications according to these criteria.

McCrocklin et al. (2019) examined the accuracy rates of Windows Speech Recognition (WSR) and Google Voice Typing in recognizing the speech of advanced non-native English speakers and found that Google's system outperformed WSR. Inceoglu et al. (2023) compared the assessments generated by Google Assistant for four non-native speakers' accented English to those of native listeners. Their findings showed that the consistency between ASR evaluations and native listener judgments varies depending on the speaker and the type of oral production.

A more recent study evaluated the accuracy of America English transcriptions produced by five speech-to-text applications: Google Docs Voice Typing, Apple Dictation, Windows 10 Dictation, Dictation.io and "Transcribe" by comparing them with human-generated transcriptions (Hirai & Kovalyova, 2024). Thirty non-native speakers completed four speaking tasks, including reading a short passage and answering freely to questions. Consistent with the findings of Inceoglu et al. (2023), the study revealed that accuracy is shaped not only by the ASR systems' recognition capabilities but also by the type of speech and the influence of learners' L1 on their L2 productions. These results suggest that different ASR systems may yield varying levels of accuracy and, consequently, differing degrees of instructional effectiveness for non-native speakers.

Despite extensive research on ASR accuracy in EFL contexts, this issue has received little attention in the field of CFL. How effectively ASR systems assess the pronunciation of CFL learners remains largely unexplored.

2.3 Pronunciation Difficulties

As noted earlier (Derwing et al., 2000), it is essential to determine whether an ASR has the potential to identify learners' production difficulties. Below, we summarize pronunciation challenges commonly observed among Japanese learners of CFL.

The pronunciation difficulties of Japanese learners primarily arise from fundamental differences between the phonological systems of Chinese and Japanese. Through contrastive analysis, Lin (2019) observed that the Chinese phonetic inventory is more complex than that of Japanese, containing retroflex consonants not found in Japanese, lacking the vowel /e/, and presenting challenges in distinguishing the nasal finals /n/ and /ng/. Using the NTNU Chinese Learner Corpus of Interlanguage Phonology, Fang et al. (2015) conducted a systematic error analysis and reported that, for initial consonants, the labiodental fricative /f/ was often realized as the Japanese bilabial aspirated sound /フ/. Errors involving aspirated consonants were also common, particularly when aspirated sounds occurred in word-final positions. Regarding final errors, the high rounded vowel /u/ showed the highest error rate among all finals. For the rounded vowel /ü/, Japanese learners tended to struggle due to the lack of rounded front vowels in Japanese and their unfamiliarity with lip rounding in this context.

These findings suggest that if an ASR system can evaluate learners' utterances in a manner comparable to human teachers, it could greatly assist teachers in providing targeted corrective feedback and enable learners to address errors promptly during CAPT activities.

In this study, we investigate how accurately three ASR systems assess the Mandarin Chinese speech of Japanese learners by comparing their assessments with those of native Chinese-speaking teachers. We address this issue by examining the following research questions:

1. To what extent do these ASR systems assess learners' word-level utterances in a manner comparable to teachers?
2. Do these ASR systems have the potential to identify pronunciation difficulties?

3. Methodology

3.1 Recognizers and Words to Pronounce

3.1.1 Automatic Speech Recognition Systems

This study employed three automatic speech recognition (ASR) systems: Whisper by OpenAI, Azure by Microsoft, and Gladia by Gladia Inc. All three systems support multiple languages and demonstrate high recognition accuracy, having been widely adopted in commercial transcription services. Whisper, in particular, is open source, providing substantial flexibility for research and development. In the present study, a downloaded desktop version of Whisper was used (Whisper Desktop, 2023), while the

other two ASR systems were accessed via their online platforms (Microsoft, 2024; Gladia, 2024).

3.1.2 Word Selection

Based on Fang et al.'s (2015) findings that Japanese learners of Mandarin primarily struggle with distinguishing aspirated and unaspirated consonants, producing the vowel /e/ (which does not exist in Japanese), articulating the labiodental fricative /f/, realizing retroflex consonants such as /zhi/, /chi/, /shi/, /ri/, and /er/, and pronouncing nasal finals including /an/, /ang/, /en/, and /eng/, this study selected 20 two-syllable words that contain these pronunciation features as speech materials for a preliminary investigation. Based on our teaching experience, we considered two-syllable words to be relatively easy for beginning learners to pronounce. The selected items were aligned with the learners' instructional progression. The target words are listed in Table 1.

Table 1 Words to Pronounce

客气, 学校, 放心, 词典, 听懂, 注意, 好吃, 老师, 日本, 二十
咖啡, 工作, 你家, 买菜, 北京, 便宜, 很远, 常常, 告诉, 别走

3.2 Scoring Criteria

3.2.1 Speech Intelligibility and Comprehensibility

Foreign language learner pronunciation assessment can be approached from the perspective of native listener comprehension, which is typically divided into two dimensions: *intelligibility* and *comprehensibility* (Munro & Derwing, 1999). *Intelligibility* refers to the extent to which listeners can accurately identify the linguistic content produced by the speaker (e.g., phonemes and words), emphasizing objective phonetic recognition. *Comprehensibility*, on the other hand, concerns the degree of effort required for listeners to understand the speaker's intended meaning, representing a more subjective, global evaluation.

In the context of ASR assessment, intelligibility constitutes the primary evaluative dimension (Inceoglu et al., 2023), as ASR systems rely predominantly on acoustic features and lack the ability to process higher-level linguistic, contextual, or pragmatic information. However, in real classroom settings, teachers often find learners' speech easier to understand than native listeners without teaching experience, particularly those unfamiliar with non-native speech patterns. Since listener comprehension can influence learners' motivation to engage in pronunciation practice, comprehensibility remains an important perspective in assessing foreign language pronunciation. Although *intelligibility* and *comprehensibility* are theoretically distinct constructs, they may be treated similarly in classroom practice for pedagogical purposes.

In other words, if speech recognizers could "understand" learner speech in a manner similar to teachers, learners might be more motivated to practice pronunciation. Therefore, this study compares ASR recognition outcomes (intelligibility) with teacher evaluations of comprehensibility to examine the degree of alignment between ASR systems and human

instructors in assessing learners' word-level Mandarin pronunciation. Through this comparison, we aim to explore the potential for ASR systems to provide teacher-like assessments.

3.2.2 Scoring Learner Speech

Table 2 Pronunciation Scoring Criteria

Score	ASR Criteria	Teacher Criteria
0	Both characters unrecognizable	Word meaning incomprehensible
1	One character correctly recognized	Pronunciation unclear but meaning comprehensible
2	Both characters correctly recognized	Pronunciation clear and meaning comprehensible

Since ASR systems output only recognition results, each recognized Chinese character was regarded as correct if it corresponded to the target character shown in Table 1. This study adopted a three-level scoring scheme for each word of the 20 two-syllable words: 0 points if neither character was recognized, 1 point if only one character was correctly recognized, and 2 points if both characters were correctly recognized. This scoring method corresponds to the Character Error Rate (CER) evaluation framework, which is calculated based on three error types—substitution, deletion, and insertion. As the focus of this study is on pronunciation at the word level, homophonic outputs generated by the ASR systems were treated as correct.

Teacher scoring was conducted by five native Mandarin-speaking instructors with extensive experience teaching Chinese as a foreign language in Japan. This suggests that the instructors are familiar with Japanese learners' pronunciation and may therefore be more tolerant when evaluating learner speech. The raters listened to the learners' recorded utterances and assigned scores on a three-point comprehensibility scale: 0 points if the word's meaning was completely incomprehensible, 1 point if the pronunciation was unclear but the meaning remained interpretable, and 2 points if the pronunciation was clear and the meaning fully comprehensible. The five teachers rated the samples independently without consultation. This multi-rater design reduces the influence of individual subjective tendencies and enhances the overall reliability of the scoring. The scoring criteria are summarized in Table 2.

3.3 Speakers and Procedure of Recording Utterances

The learner speech data consisted of word-level utterances produced by 31 university students who were taking introductory Chinese courses for the first time. These students received two 90-minute Chinese classes per week and had completed phonetic instruction prior to the recording sessions. The target words, along with their pinyin and Japanese translations, were provided for in-class pronunciation practice, and additional practice was assigned as homework.

Students were instructed to make their recordings in quiet environments using their smartphones and to submit the audio files via the learning management system at the university. They made the recordings individually. The researcher then converted the submitted audio files into formats compatible with each ASR system and scored each learner's pronunciation of each word based on the criteria shown in Table 2.

4. Results

4.1 Interrater Reliability

To address RQ1, we summed the scores of all target words rated by the three ASR systems and the five teachers for each student. Each total score ranged from 0 to 40. Although the data were quantitative, they did not follow a normal distribution.

Before using the mean teacher scores as the representative measure of human assessment, we first examined interrater reliability among the five teachers. The intraclass correlation coefficient (ICC) was calculated based on the total scores, and Fleiss's kappa coefficients were computed for each individual word. All statistical analyses were conducted using IBM SPSS Statistics 30.0.

The results indicated substantial agreement among the teachers (ICC = 0.63) (Landis & Koch, 1977). Across the 20 words, Fleiss's kappa values ranged from 0.20 (slight agreement) to 0.63 (substantial agreement). Specifically, one word showed slight agreement (北京: $\kappa = 0.20$, $z = 3.69$, $p < .001$), 11 words showed fair agreement ($\kappa = 0.21$–0.40), seven words showed moderate agreement ($\kappa = 0.41$–0.60), and one word showed substantial agreement (别走: $\kappa = 0.64$, $z = 11.93$, $p < .001$). Among the ASR systems, a fair level of agreement was observed (ICC = 0.55). Following Inceoglu et al. (2023), we consider the teacher ratings to be sufficiently reliable and therefore use the mean teacher scores as the representative measure of teacher assessment.

4.2 Rating Results

The values shown in Table 3 represent the mean scores and standard deviations of students' total scores as assessed by the three ASR systems and the teachers. To maintain consistency with the scoring criteria described in Section 3.2.2, the total scores were divided by 20 so that the resulting values correspond to the mean and standard deviation per word. T1–T5 denote the five teachers.

Table 3 Rating Results by ASR and teachers

Rater	Whisper	Azure	Gladia	T1	T2	T3	T4	T5	Teacher
M	1.46	1.25	1.42	1.69	1.56	1.57	1.45	1.38	1.53
SD	0.27	0.30	0.28	0.17	0.24	0.24	0.34	0.32	0.24

The results show that the mean teacher score was 1.53 (SD = 0.24). Among the three ASR systems, Whisper achieved the highest performance (M = 1.46, SD = 0.27),

followed by Gladia (M = 1.42, SD = 0.28), while Azure demonstrated the lowest performance (M = 1.25, SD = 0.30).

The individual teacher scores (T1–T5) indicate that, although the overall standard deviation (0.24) was lower than those of the ASR systems, noticeable variation remained across teachers, with standard deviations ranging from 0.17 (T1) to 0.34 (T4). These findings suggest that even among experienced Chinese language instructors, individual differences persist in evaluating learner pronunciation. This underscores the challenge of establishing fully standardized human assessment and highlights the potential utility of ASR-based evaluation.

The Wilcoxon signed-rank test results indicated that the difference between Whisper scores and teacher scores was not statistically significant, whereas the scores from Azure and Gladia differed significantly from teacher scores. These findings suggest that Whisper provided a more teacher-like assessment of the 20-word utterances produced by the 31 learners compared with Azure and Gladia.

To present the differences between ASR systems and teacher ratings in a more intuitive manner, we calculated percentage difference scores using the formula

(ASR Score − Teacher Score) / Teacher Score × 100%.

The results are summarized in Table 4.

Table 4 Percentage Differences

ASR System	Percentage Difference
Whisper	-4.32%
Azure	-18.57%
Gladia	-7.09%

On average, the results suggest that Azure had the most difficulty recognizing the learners' Chinese word-level utterances compared with the other two systems.

4.3 Influence of Learner Proficiency

Based on the five-number summary of the teacher average scores, the 31 students were divided into three proficiency groups:

1. Low-proficiency group (0.99–1.45): 10 learners (Learners 1–10)
2. Medium-proficiency group (1.48–1.58): 10 learners (Learners 11–20)
3. High-proficiency group (1.64–1.96): 11 learners (Learners 21–31)

Learner 1 received the lowest teacher score (0.99), whereas Learner 31 received the highest (1.96). In comparison, the ASR system ratings yielded the following distributions:

- Whisper classified 5 learners as low-, 4 as medium-, and 10 as high-proficiency;

- Azure classified 6 learners as low-, 4 as medium-, and 7 as high-proficiency;
- Gladia classified 7 learners as low-, 5 as medium-, and 8 as high-proficiency.

These results suggest that the consistency between ASR systems and teacher assessments increased for learners in the high-proficiency group.

Regarding extreme values, Whisper assigned the lowest score to Learner 4 (0.60), Azure to Learners 3 and 16 (0.70), and Gladia to Learner 5 (0.75). Whisper's highest score was for Learner 22 (1.90), Azure's for Learner 27 (1.85), and Gladia's for Learners 22, 24, and 28 (1.80). The variation in highest and lowest scores across teachers and ASR systems indicates that ASR assessments are influenced by learner proficiency levels.

4.4 Extremum Cases and Pronunciation Characteristics

According to the percentage difference values, the ASR scores for Learner 1 were as follows: Whisper overestimated the learner's performance by 36% (1.35), Azure underestimated it by 14% (0.85), and Gladia overestimated it by 46% (1.45).

Several characteristics emerged from Learner 1's pronunciation analysis conducted by the authors. The pronunciations of "老师" and "日本" were clearly inaccurate, and most teachers assigned scores of 0 or 1. However, both Whisper and Gladia correctly recognized these utterances. Conversely, Whisper and Azure failed to recognize "客气", whereas Gladia succeeded. These findings suggest that ASR assessments for low-proficiency learners are influenced not only by the learners' pronunciation characteristics but also by system-specific recognition criteria In this case, Azure's score was the closest to the teacher assessment, and the difference was not statistically significant according to the Wilcoxon signed-rank test.

For the highest-proficiency learner (Learner 31, teacher score = 1.96), the ASR scores were as follows: Whisper underestimated the performance by 18% (1.61), Azure by 13.27% (1.70), and Gladia by 11% (1.74). Pronunciation checks conducted by the authors revealed that Learner 31's pronunciations were highly native-like and contained no noticeable segmental errors. However, the duration of each pronunciation was relatively long, which may have caused recognition difficulties for the ASR systems. Sunaoka (2018) notes that excessive pronunciation length can negatively affect recognition accuracy, potentially leading to insertions, omissions, and other errors. In this case, Gladia's assessment was the closest to the teacher evaluation, and the difference was not statistically significant based on the Wilcoxon signed-rank test.

5. Analysis of Pronunciation Characteristics

5.1 Rating Results of Six Pronunciation Categories

To address RQ2—whether these recognizers have the potential to identify pronunciation difficulties—we examined the ASR ratings from the perspective of pronunciation characteristics. The 20 target words were classified into six major categories

based on the initial consonant of the first character: bilabial, apical, velar, palatal, retroflex, and alveolar. Because of the small sample size in each category, the following findings should be regarded as exploratory and interpreted as case-study evidence. Table 5 presents the mean ASR percentage difference scores for each category.

Table 5 ASR Percentage Difference Scores Based on Pronunciation Categories

Pronunciation Category	Whisper	Azure	Gladia
Bilabial	1.9%	-17.0%	5.0%
Apical	-1.8%	-10.7%	-7.7%
Velar	-9.4%	-23.3%	-12.6%
Palatal	17.6%	5.7%	17.6%
Alveolar	-34.2%	-53.0%	-17.1%
Retroflex	-9.8%	-18.0%	-23.3%

The results indicate that ASR ratings varied across pronunciation categories. All ASR systems tended to underestimate learner performance relative to teacher ratings, with the exception of palatal sounds. For apical sounds, the percentage differences across the systems were relatively small, whereas alveolar sounds exhibited much larger variation. The mean teacher score for apical sounds was1.69, while alveolar sounds received a lowerr mean score of1.17. These findings suggest that the degree of consistency between ASR assessments and teacher evaluations is strongly influenced by the specific pronunciation characteristics of each sound category.

5.2 Bilabial Sound Words

Table 6 ASR and Teacher Scores of Bilabial Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
便宜	p bilabial/zero initial	ian/i	1.74	1.03	1.74	1.37
别走	b bilabial/z alveolar	ei/ou	1.48	1.45	1.55	1.64
北京	b bilabial/j palatal	ei/ing	1.87	1.65	1.87	1.75
买菜	m bilabial/c alveolar	ai/ai	1.74	0.97	1.74	1.72
放心	f labiodental/x palatal	ang/in	1.26	1.48	1.45	1.45

Table 6 presents the scores of the bilabial sound words assigned by the ASR systems and the teachers. The results of the Wilcoxon signed-rank tests indicate significant differences between teacher ratings and ASR scores for “便宜” (Whisper: $p = .003$; Gladia: $p = .045$), “北京” (Whisper: $p = .028$; Gladia: $p = .020$), and “买菜” (Azure: $p < .001$), whereas no significant differences were observed for “别走” or “放心.”

For “便宜,” which involves the aspirated consonant /p/ combined with the compound final /ian/, Whisper and Gladia assigned higher scores than the teachers. Azure’s ratings did not significantly differ from teacher assessments. Student pronunciation

analysis conducted by the authors confirmed that while learners could distinguish between aspirated /p/ and unaspirated /b/, both the strength and duration of aspiration were insufficient, which may have contributed to the different recognition outcomes across systems.

For “买菜,” Azure assigned an exceptionally low score, with recognition errors frequently occurring in the second syllable—for example, producing “在” or “开.” Learner pronunciation checks revealed that although students’ /c/ pronunciations were distinguishable, insufficient aspiration strength may have caused Azure's performance to diverge from that of the other two ASR systems.

Although Japanese lacks the /f/ sound and previous research by Fang et al. (2015) reported that /f/ is often confused with /h/ by Japanese learners, such confusion was not observed for “放心” in this study. A possible explanation is that the vowel /a/ exists in Japanese, and when combined with /f/, the overall phonetic structure becomes easier for learners to produce, reducing the likelihood of confusion.

5.3 Apical Sound Words

Table 7 ASR and Teacher Scores of Apical Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
听懂	t apical/d apical	ing/ong	1.48	1.29	1.58	1.72
你家	n apical/j palatal	i/ia	1.74	1.65	1.26	1.68
老师	l apical/sh retroflex	ao/zero	1.74	1.58	1.84	1.68

There were three apical-initial words in the learner dataset, as shown in Table 7. The results of the Wilcoxon signed-rank tests indicate that a significant difference between ASR and teacher scores occurred only for “老师” (Gladia: $p = .023$). This suggests that Gladia may overestimate learners’ productions of apical sounds, whereas the assessments provided by Whisper and Azure were consistent with teacher ratings.

5.4 Velar Sound Words

Although six words contained velar-initial sounds, the Wilcoxon signed-rank test results showed no significant differences between ASR and teacher scores for five of the items. The only exception was “客气”, for which significant differences were found across all systems (Whisper: $p < .001$; Azure: $p < .001$; Gladia: $p = .003$).

Table 8 ASR and Teacher Scores of Velar Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
客气	k velar/q palatal	e/i	0.13	0.16	0.52	1.06
告诉	g velar/s alveolar	ao/u	1.45	1.32	1.26	1.60
工作	g velar/z alveolar	ong/uo	1.94	1.48	1.94	1.79

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
很远	h velar/zero initial	en/yuan	1.71	1.23	1.29	1.57
好吃	h velar/ch retroflex	ao/zero	1.64	1.55	1.39	1.52
咖啡	k velar/f bilabial	a/ei	1.90	1.87	1.74	1.84

The ratings for "客气" indicate that students' primary difficulties with velar sounds involve the /e/ final and aspiration control. Learner pronunciation checks showed that only a few of the 31 students produced "客气" with relative accuracy. The pronunciation errors observed fell into three main categories: (1) insufficient aspiration of /k/, (2) errors in the /e/ final, and (3) tone errors. ASR recognition outputs included forms such as "各级," "各起," and "课题," reflecting these deviations.

In contrast, ASR recognition of "咖啡," which shares the same initial /k/, exhibited high consistency with teacher ratings. This disparity suggests that ASR recognition errors for velar-initial words vary according to initial–final combinations. In other words, such contrasts may help elucidate specific pronunciation difficulties among learners.

5.5 Palatal Sound Word and Alveolar Sound Word

Table 9 ASR and Teacher Scores of Palatal Sound and Alveolar Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
学校	x palatal/x palatal	ue/iao	1.87	1.68	1.87	1.59
词典	c alveolar/d apical	i/ian	0.77	0.55	0.97	1.17

Only one palatal-initial word and one alveolar-initial word were included in the learner data. The Wilcoxon signed-rank test results showed significant differences between teacher and ASR scores for "学校" (Whisper: $p = .018$; Gladia: $p = .010$) and for "词典" (Whisper: $p = .002$; Azure: $p < .001$).

For the palatal-initial word "学校," ASR systems tended to overestimate learner performance. Learner pronunciation checks indicated that students generally produced the palatal fricative /x/ accurately, whereas the final /u/ was often omitted or realized as "xie." These deviations may have been insufficiently penalized by the ASR systems, leading to higher scores than those assigned by teachers.

For the alveolar-initial word "词典," ASR systems tended to underestimate performance relative to teacher ratings. Recognition outputs frequently included "ji dian"–type errors (e.g., "寄典," "机点"), suggesting that learners often substituted the aspirated alveolar affricate /c/ with the unaspirated palatal affricate /j/. This substitution reflects insufficient aspiration in producing /c/, and ASR systems appeared highly sensitive to this cue, resulting in lower scores.

5.6 Retroflex Sound Words

Retroflex consonants constitute a phonetic category unique to Mandarin and are entirely absent from the Japanese phonological system; thus, they represent one of the greatest pronunciation challenges for Japanese learners. Both ASR and teacher ratings for retroflex-initial words tended to be low. The Wilcoxon signed-rank test showed a significant difference only for Gladia's score for "二十" ($p < .001$).

Learner pronunciation checks indicated that the primary difficulty with "常常" stemmed from the retroflex affricate /ch/. Many students failed to produce sufficient tongue-tip retroflexion, resulting in ASR outputs such as "江" and "强." Additionally, because "二十" contains the retroflex final /er/, this likely contributed to Gladia's substantially lower score relative to teacher ratings.

Table 10 ASR and Teacher Scores of Retroflex Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
常常	ch retroflex/ch retroflex	ang/ang	1.10	0.65	0.71	1.10
注意	zh retroflex/zero initial	u/i	1.16	0.84	1.61	1.21
二十	zero initial/sh retroflex	er/zero	1.10	1.55	0.23	1.41
日本	r retroflex/b bilabial	zero/en	1.45	1.32	1.55	1.59

Overall, the word-level analyses preliminary demonstrate that ASR assessments are strongly affected by specific pronunciation characteristics, and that initial–final combinations play a crucial role in determining the degree of alignment between ASR and teacher evaluations. Individual initials or finals may yield high recognition accuracy when paired with certain syllables, as in "咖啡," or low accuracy when combined differently, as in "客气." Such contrasts in acoustic features suggest that ASR outputs may help identify learner pronunciation difficulties based on systematic patterns across syllable structures.

6. Conclusion and Future Work

In this study, we examined how effectively three ASR systems—Whisper, Azure, and Gladia—evaluate Japanese learners' Mandarin word-level pronunciation by comparing ASR-generated scores with ratings provided by experienced teachers. Two research questions were addressed: (1) To what extent do ASR systems assess learner pronunciation in a manner consistent with teachers? (2) Do ASR systems have the potential to identify learner pronunciation difficulties? A scoring scheme was developed to quantitatively evaluate the utterances of 31 Japanese learners producing 20 Chinese words, and ASR–teacher comparisons were conducted from the perspectives of learner proficiency and pronunciation characteristics.

Analyses based on learner-level scores showed that although all three ASR systems tended to underestimate learner performance relative to teachers, Whisper provided the most teacher-like assessments overall. With regard to individual proficiency levels, Azure

aligned most closely with teacher ratings for the lowest-proficiency learner, whereas Gladia showed the highest alignment for the highest-proficiency learner. These patterns suggest that ASR performance is influenced by learner proficiency. These results are partly consistent with those reported by Hirai and Kovalyova (2024), who also observed variability in ASR performance depending on phonetic features for English non-native speakers.

Word-level analyses preliminary revealed that ASR assessments were strongly affected by specific pronunciation features, particularly the interaction between initials and finals. Velar-initial words showed the highest overall consistency with teacher ratings, except for "客气," whereas alveolar-initial words exhibited the largest discrepancies. Single-syllable initials or finals may yield high or low recognition accuracy depending on their syllabic combination, as demonstrated by the contrast between "咖啡" and "客气." These contrasts indicate that ASR outputs may be used to identify pronunciation difficulties by analyzing systematic acoustic deviations.

The findings suggest that ASR systems have the potential to provide teacher-like assessments of learners' pronunciation. However, learner proficiency and pronunciation features should both be taken into account when implementing ASR systems for pronunciation learning and practice.

Additionally, the standard deviations among the five professional teachers ranged from 0.17 to 0.34, demonstrating that even trained raters exhibit individual differences when assessing learner pronunciation. This finding underscores the need for objective scoring methods and highlights the potential value of ASR systems as auxiliary tools in computer-assisted pronunciation training (CAPT). While ASR systems cannot fully replace human evaluators, they can provide consistent, phoneme-level assessment and reduce teacher workload, particularly in large-scale or formative assessment contexts.

This study has several limitations. Future research should expand the word set to encompass a wider range of phonetic features and include learners with more diverse proficiency levels. Further work should also examine ASR performance at the sentence and paragraph levels and incorporate additional scoring criteria to support the development of more comprehensive and objective assessment frameworks.

In summary, the findings highlight both the potential and the current limitations of applying ASR technologies to CAPT for Mandarin Chinese and provide insights for designing more effective pronunciation assessment and feedback systems.

References

Burleson, D. F. (2007). Improving intelligibility of non-native speech with computer-assisted phonological training. *Indiana University Linguistics Club Working Papers, 7*(1), 1-18. https://scholarworks.iu.edu/journals/index.php/iulcwp/article/view/25801

Chen, J. (2011). Application of VoiceThread in Chinese teaching and learning: Some examples. *Journal of Technology and Chinese Language Teaching, 2*, 81-94. [蘇芳儀. (2011). VoiceThread 應用於中文教學的幾個例子. *科技与中文教学, 2*(1), 81-94.] http://www.tclt.us/journal/2011v2n1/chenj.pdf

Da, J. (2015). The application of speech recognition technology in Chinese language learning: What can be learned from a Pinyin lab session. *Journal of Technology and Chinese Language Teaching, 6*(1), 16-24. [笪骏. (2015). 语音识别技术在中文教学中的应用：一堂汉语拼音练习课的启示. *科技与中文教学, 6*, 16-24.] http://www.tclt.us/journal/2015v6n1/da.pdf

Dai, Y. & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning, 36*, 861-884. https://doi.org/10.1080/09588221.2021.1952272

Derwing, T. M., Munro, M. J. & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34(3)*, 592-603. https://doi.org/10.2307/3587748

Ehsani, F. &Knodt, E. (1998). Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm. *Language Learning & Technology*, *2*(1), 54-73. https://doi.org/10.64152/10125/25032

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832–844. https://doi.org/10.1016/j.specom.2009.04.005

Fang, S., Chen, C., Wang, C., Yang, H., & Chen, H. (2015). An error analysis on Japanese learners' Chinese pronunciation with the aid of Chinese learners' oral corpus. *Journal of Chinese Language Teaching, 12*(3), 93-123. [方淑華, 陳慶華, 王敬淳, 楊惠媚, &陳浩然. (2015). 藉學習者口語語料庫探究日籍生常見的華語語音偏誤與教學建議. 華語文教學研究, *12*(3), 93-123.]

Fouz-González, J. (2015). Trends and directions in computer-assisted pronunciation training. In J. A. Mompean (Ed.), *Investigating English pronunciation: Trends and directions* (pp. 314–342). Palgrave Macmillan. https://doi.org/10.1057/9781137509437_14

Gao, F. (2025). A study on the learning motivations, goals, difficulties, and expectations of Japanese university students learning Chinese as a foreign language. *Journal of Aichi Shukutoku University*, *15*, 49-62. [高飛. (2025). 外国語として中国語を学習している日本人大学生の学習動機、目的、困難点及び期待. *愛知淑徳大学論集—交流文化学部篇— 第15 号*, 49-62.] https://aska-r.repo.nii.ac.jp/records/2000370

Gladia. (2024). *Gladia* [Audio transcription software]. https://www.gladia.io/

Hirai, A., & Kovalyova, A. (2024). Speech-to-text applications' accuracy in English language learners' speech transcription. *Language Learning & Technology, 28*(1), 1–21. https://doi.org/10.64152/10125/73555

Inceoglu, S., Chen, W. H. & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL, 35*(1), 89–104. https://doi.org/10.1017/S0958344022000192

Issa, E., & Hahn-Powell, G. (2025). Computer-assisted pronunciation training for foreign language learning of grammatical features. *Language Learning & Technology, 29*, 1–20. https://doi.org/10.64152/10125/73622

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Li, N., Zhang, L., Lau, K. L., & Liang, Y. (2024). Predicting Chinese language learners' ChatGPT acceptance in oral language practices: The role of learning motivation and willingness to communicate. *Journal of Technology and Chinese Language Teaching, 15*(1), 25-48. [李诺恩, 张岚, 刘洁玲, & 梁宇. (2024). 预测中文学习者在口语练习中对ChatGPT的接受度：学习动机和交流意愿的作用. *科技与中文教学, 15*, 25-48.] http://www.tclt.us/journal/2024v15n1/lizhanglauliang.pdf

Lin, C. (2019). The phonetic problems of Japanese Chinese language learners and teaching suggestions. *TCSL Forum,* 27, 9-37. [林嘉惠. (2019). 日籍華語學習者的語音問題與其教學建議. *華語學刊*, 27, 9-37.] https://www.airitilibrary.com/Article/Detail/P20151202001-201912-202005290011-202005290011-9-37

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? *In J. Levis, C. Nagle, & E. Todey (Eds.), Proceedings of the 10th pronunciation in second language learning and teaching conference* (pp. 191–200), Iowa State University. https://www.iastatedigitalpress.com/psllt/article/id/15376/

Microsoft. (2024). *Azure AI Speech* [Audio transcription software]. https://azure.microsoft.com/products/ai-services/ai-speech

Munro, M. J. & Derwing, T. M. (1999) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*(s1), 285-310. https://doi.org/10.1111/0023-8333.49.s1.8

Neri, A., Cucchiarini, C. & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL, 20*(2), 225-243. https://doi.org/10.1017/S0958344008000724

O'Brien, M. G, Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I, Strik, H., Levis, J. M., Munro, M. J., Foote, J. A. &Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4*(2), 182 – 207. https://doi.org/10.1075/jslp.17001.obr

Sunaoka, K. (2018). Using automatic speech recognition technology to reverse analyze communication strategies between non-native speakers in a Chinese long distance group discussion. *Journal of Technology and Chinese Language Teaching, 9*(2), 61-82. [砂冈和子. (2018). 以语音识别技术逆向分析汉语远场群体讨论中非母语者的交互策略. 科技与中文教学, *9*(2), 61-82.] http://www.tclt.us/journal/2018v9n2/sunaoka.pdf

Tsai, P. (2019). Beyond self-directed computer-assisted pronunciation learning: A qualitative investigation of a collaborative approach. *Computer Assisted*

Language Learning, 32(7), 713-744. https://doi.org/10.1080/09588221.2019.1614069

Wang, Y. & Young, S. S. (2014). A study of the design and implementation of the ASR-based iCASL system with corrective feedback to facilitate English learning. *Educational Technology & Society, 17*(2), 219-233. https://www.jstor.org/stable/jeductechsoci.17.2.219

Watanabe, Y., Omae, T. & Odo, S. (2019). Investigating the effect of chinese pronunciation teaching materials using speech recognition and synthesis functions. *Journal of Technology and Chinese Language Teaching*, *10*(2), 102-124. http://www.tclt.us/journal/2019v10n2/watanabeomaeodo.pdf

Wei, W. & Zhang, J. (2018). An intelligent Chinese pronunciation teaching app and the preliminary result of a teaching experiment. *Journal of Technology and Chinese Language Teaching*, *9*(2), 83-97. [魏巍 & 张劲松. (2018). 一款汉语智能语音教学App及教学实验初步结果. *科技与中文教学, 9*(2), 83-97.] http://www.tclt.us/journal/2018v9n2/weizhang.pdf

Whisper Desktop. (2023). *Whisper* (small version 17.4.4.) [Audio transcription software]. https://github.com/const-me/whisper

Zhao, X., Tomita, N., Konno, F., Ohkawa, Y. & Mitsuishi, T. (2019). Development and practice of review material KoToToMo for use on smartphones in blended learning by beginning learners of Chinese in university. *Transactions of Japanese Society for Information and Systems in Education, 36*, 131-142. [趙秀敏，冨田昇，今野　文子, 大河雄一 & 三石大. (2019). 大学初修中国語ブレンディッドラーニングのためのスマートフォン利用復習教材「KoToToMo」の開発と実践. *教育システム情報学会誌. 36*, 131-142.] https://www.jstage.jst.go.jp/article/jsise/36/2/36_360211/_article/-char/ja/

