TalkPro: A Multimodal Language Learning and Evaluation System

Johannes Wirth and René Peinl
Institute for Information Systems at Hof University, Hof, Germany

Johannes.wirth.3@iisys.de
Rene.peinl@iisys.de

Abstract: As universities around the world welcome increasing numbers of international students, there is a growing demand for scalable, objective tools that can support both language learning and applicant selection based on spoken language proficiency. In particular, pronunciation and comprehension remain persistent challenges for non-native speakers and are key factors for communication in academic environments. Traditional methods of assessing these skills are labor-intensive or often rely on surface-level metrics such as transcription accuracy, which do not fully capture a learner's communicative competence. This work introduces TalkPro, a multi-modal system for pronunciation and comprehension assessment as well as language learning, designed to address this need. The system provides continuous, personalized feedback on learners' spoken language, with a specific focus on phoneme-level accuracy as well as articulatory patterns. Instead of relying solely on conventional speech recognition outputs, which are often able to compensate even major pronunciation errors, TalkPro generates detailed acoustic analyses that pinpoint learner-specific difficulties. These include not only phoneme-level errors but also recurring articulatory tendencies, such as misplacement of the tongue, incorrect voicing, or inappropriate manner of articulation. The system also includes a text-to-speech (TTS) engine to generate spoken content adapted to vocabulary gaps, which is then followed by targeted comprehension questions. TTS can also be used to test listening comprehension either word by word in a dictation style or semantically using a large language model (LLM) as a judge. Overall, these components form an integral approach to pronunciation and comprehension training in a blended learning environment and can also be used for automated assessment. Preliminary experiments with incoming students from India to Germany indicate that phoneme-level ASR effectively identifies pronunciation errors, whereas grapheme-level ASR tends to overlook them. Future research will involve a comprehensive evaluation of automated results against human judgment, alongside the expansion of TalkPro's training capabilities with LLM-based reading comprehension modules that prioritize conceptual understanding over traditional verbatim recall.

Keywords: Computer-Aided Pronunciation Training, Second Language (L2), Language Learning, Speech Recognition, Text-to-speech, Large Language Model

1. Introduction

Despite the growing diversity of international students in universities worldwide, tools for assessing and developing spoken language skills in L2 (second language) learners remain underdeveloped. Most automated systems without human expert feedback were not originally developed for and thus are limited in their ability to assist in language learning (Liu et al., 2025), as they often ignore semantics, fail to align with human ratings (Kim et al., 2022) or cannot capture critical phonetic issues: voicing errors, tongue placement, and manner of articulation, that affect intelligibility in spoken communication. Meanwhile, commercial platforms for language learning like Duolingo offer only basic binary pronunciation feedback, which is insensitive to subtle mispronunciations and inconsistently applied. As a result, learners typically make stronger gains in reading and writing, while speaking and listening lag behind (Loewen et al., 2019). These gaps in both assessment and training tools for speaking and listening have concrete consequences: learners may struggle to participate in classroom discussions, lack confidence in presentations, and face communication barriers with native speakers. In today's globalized labor market, where international mobility and global employment opportunities are increasingly important, such limitations hinder academic progression and career prospects.

To bridge these gaps, this work introduces TalkPro, a multimodal platform specifically designed to improve both the assessment and training of spoken language by addressing the limitations outlined above. TalkPro integrates advanced acoustic and articulatory analysis, adaptive content generation, and intelligent comprehension exercises, built on multilingual phoneme- as well as grapheme-based automatic speech recognition (ASR) models, to deliver detailed, actionable feedback on pronunciation and understanding. By focusing on phoneme-level alignment and articulatory patterns, TalkPro identifies the specific challenges learners face, enabling targeted improvement. Its integration of Text-to-Speech (TTS) technology allows for personalized practice and correctional measures regarding pronunciation, while LLM-driven assessment (Large Language Model, LLM) ensures learners develop both accuracy and contextual understanding based on their personal weaknesses. By moving beyond surface-level metrics and focusing on the nuanced aspects of pronunciation and comprehension, TalkPro represents a significant step toward equitable, scalable, and effective language learning as well as assessment.

2. Literature Review

Between 1990 and 2010, CLT emphasized meaning over form (Pennington 2021), but in the last decade pronunciation regained importance with greater acceptance of variation. The primary goal is intelligibility and effective communication, rather than training towards a single, standardized model of perfect pronunciation (ibid). Although it is possible for adults to learn accurate pronunciation of a foreign language according to Flege's Speech Learning Model (SLM) by reorganization to allow for L2 sounds and add new phonetic categories, this is not an easy or quick process, it requires a high quality and quantity of L2 input, as well as sufficient training possibilities for L2 use in communication (Fletcher 1995).

According to Neri et al. (2008), computer-assisted pronunciation training leads to improvements in language learning for 11-year-old children that are on par with a comparison group, that was instructed by a teacher. The study focused on Italian pupils' English word-level pronunciation, using a simple ASR module limited to single-word recognition. The small sample (13 with computers, 15 without) prevented statistically significant results. Other sources like (Lee et al., 2015) suggest that instruction provided by a teacher is more effective than instruction provided on computer.

Pennington and Rogerson-Revell state in 2018, that "limitations of artificial intelligence and speech recognition and synthesis mean that it is hard for a computer to really communicate or negotiate meaning with a human". In 2025, however, this is not the case anymore. The difficulties of ASR in 2019 were dealing with accented,

non-native speech since such variations in pronunciation could not easily be represented in ASR databases. Our phoneme-based ASR highlights deviations directly, unlike earlier systems such as 'Pronunciation Power,' which relied on waveform comparison to a reference speaker (Pennington & Rogerson-Revell 2018).

CAPT systems can be tuned either in favor of falsely accepting incorrect responses (false positives) or falsely rejecting correct ones (false negatives). Since the amount of corrective feedback should be limited to avoid demotivating learners and to prioritize feedback on sounds with a high functional load (Neri et al. 2002), avoiding false negatives should be a priority. Phoneme-based ASR can be combined with grapheme-based ASR in order to prioritize pronunciation errors found with P-ASR according to their relevance for intelligibility judged by G-ASR, as suggested by Liu et al. (2025).

Bashori et al. (2024) investigated the effects of two ASR-based language learning systems, namely ILI and NOVO, on learners' word-level and sentence-level pronunciation of English language. 117 Indonesian high-school students participated in a five-week-long study (52 students used ILI and 65 NOVO). Phonetic edit distances and accentedness as well as comprehensibility ratings were calculated on a pre- and post-reading test. Results showed significant improvements in learners' pronunciation, with NOVO leading to more progress.

Liu et al. (2025) perform a systematic review of ASR use in EFL education and find that the majority of studies adopted a quasi-experimental design, primarily focusing on the pronunciation gains in accuracy. Their findings indicate that ASR is an effective tool in language classrooms but still entails limitations.

Practicing with ASR alone would not be feasible for every learner as some still require assistance to correct pronunciation mistakes (Liu et al. 2025). Therefore, we propose direct feedback from TTS to provide that kind of assistance.

Sentürk (2023) experimented with ChatGPT as a textual help for learning German. He used it for getting explanations of language constructs, example sentences and translations and found it rather helpful, although he found small mistakes in the translation in some cases.

The closest work to ours is (Moxon, 2024), which evaluates a similarly comprehensive computer system for language learning, but again for English language learning. Moxon uses Text-to-Speech, Automatic Speech Recognition, Automatic Pronunciation Assessment, and immediate visual feedback mechanisms from Microsoft Azure services to form "All-Talk". They included instructional videos showing a cut-through view of the mouth while speaking. Although changes in fluency between the pre-test and post-test were not statistically significant, male students improved significantly in overall pronunciation accuracy.

Summed up, computer-based language learning software should (Pennington & Rogerson-Revell 2019)

- be consistent with human feedback;
- be immediate;
- be pertinent and correct;

- be given in a form that students can make use of;
- include information about when goals have been reached
- suggest ways to address errors

3. The Proposed System

3.1 Overview

Our proposed system is Web-based and consists of a phoneme-level ASR, a grapheme-level ASR, a multi-speaker TTS system as well as a large language model (see Figure 1), in order to provide a comprehensive computer-based language learning offering for L2 speakers learning German. The system called TalkPro features both training and assessment functionality for speaking, listening and understanding. Users can listen to spoken text (TTS) and transcribe it to train both listening and writing or choose to provide content-equal text in either German or English (if they are proficient English speakers). Our main target group is bachelor graduates from India that come to Germany for a master study program. The LLM is judging the competence level based on ground truth and given transcript or translation on a 0 to 10 scale. The other way around, users can read German text and get feedback from both ASR systems regarding intelligibility (G-ASR) and pronunciation (P-ASR). Words that are pronounced incorrectly can be replayed with different TTS voices. Differences between recognized and expected pronunciation are highlighted in an International Phonetic Alphabet (IPA) representation. Other words with the same phonemes are displayed in order to show similarities and promote a deeper level of insight.

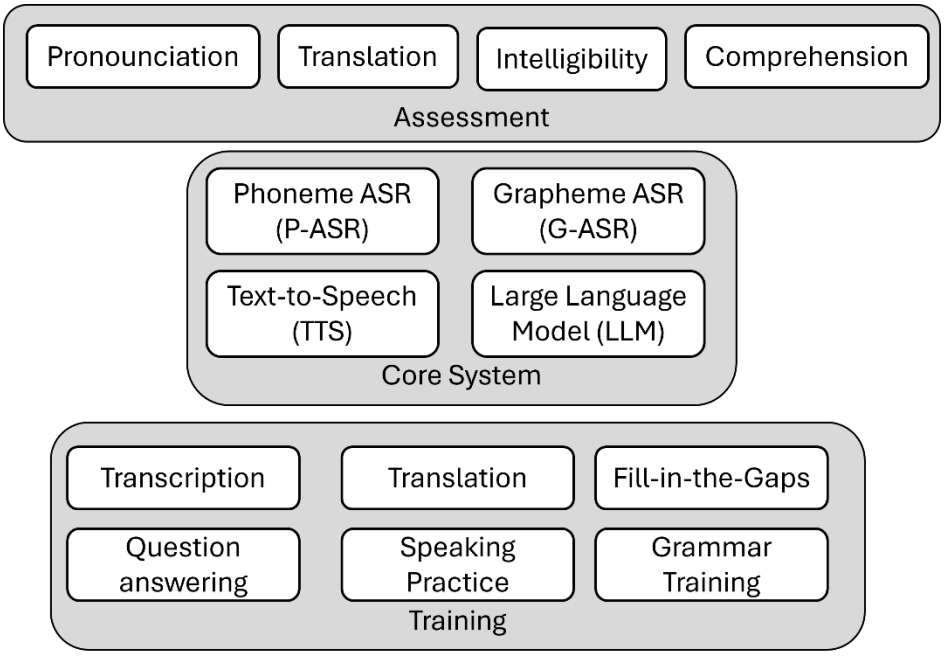

Figure 1: High-Level System Overview including training tasks, assessment possibilities and Modules.

Furthermore, translation on a textual level can be practiced as well as question answering to train text comprehension based on a short paragraph of text in German language (either written or spoken) and questions invented by the LLM.

3.2 ASR Modules

TalkPro incorporates two distinct ASR modules to evaluate and support learners' spoken German: a **phoneme-level ASR (P-ASR)** for pronunciation feedback, and a **grapheme-level ASR (G-ASR)** for intelligibility scoring. These modules enable the system to assess a user's spoken input, highlight errors, and guide improvement in pronunciation, fluency, and speech clarity.

Learners can read German text aloud, and the ASR modules analyze their speech to provide rich, multi-layered feedback:

- The **P-ASR** system performs **phoneme-level alignment and scoring**, identifying mispronunciations and mismatches between the spoken and expected phoneme sequences. Detected issues are

displayed using **IPA notation**, and incorrectly pronounced words can be replayed with alternative TTS voices for comparison.
- The **G-ASR** module estimates **intelligibility**, whether a typical listener would correctly understand the spoken input. It compares the recognized text against the ground truth and supports a **fluency-based scoring mechanism** integrated into the larger assessment workflow.

From a technical perspective, the P-ASR module is based on a Fastconformer-Transducer architecture (Gulati et al., 2020) and trained on a phonemized multilingual corpus comprising approximately 10,000 hours of speech from Common Voice (Ardila et al., 2019), Multilingual LibriSpeech (Pratap et al., 2020), and VoxPopuli (Wang et al., 2021), with data uniformly distributed across English, French, German, and Spanish. Phonemization was performed using OLaPh (Wirth, 2025), a phonemization framework that leverages Wiktionary-derived pronunciation dictionaries, statistical subword modeling based on large text corpora from Wikipedia dumps, and language-aware NLP preprocessing using spaCy (Honnibal et al., 2020). This allows the framework to handle lexical ambiguities (e.g., "wound" as noun vs. verb) and align phonemes appropriately based on sentence context.

For intelligibility estimation, the system uses Canary 1B Flash by Nvidia (Puvvada et al., 2024), a state-of-the-art grapheme-level ASR model, which complements the P-ASR by providing robust transcription and noise tolerance.

3.3 TTS Module

The TTS module provides **auditory input and corrective feedback** that supports both **listening comprehension** and **pronunciation training** for learners of German. It enables interaction with natural-sounding synthetic voices and plays a central role in exercises targeting listening, transcription, and pronunciation modeling.

Learners interact with the TTS module in two primary contexts:

- The system uses TTS to synthesize short text passages that learners transcribe. This supports auditory discrimination, spelling, and familiarity with German prosody. Passages can be predefined or generated by the LLM according to learner level.
- In pronunciation correction workflows, TTS replays the correct version of words or phrases detected as mispronounced by the ASR modules. Learners can choose to hear variations across multiple synthetic voices. Additionally, the system may synthesize phonetically similar words to highlight contrastive sounds and help refine pronunciation.

The component currently integrates five single speaker VITS (Kim et al., 2021) models based on the German datasets HUI (Puchtler et al., 2021) as well as Thorsten Voice (Müller & Kreutz, 2022). These voices vary in speaker characteristics such as gender, tone and speaking rate, enhancing the diversity of input and enabling a richer learning experience.

3.4 LLM Module

The LLM module serves as the system's **core reasoning and evaluation component**, enabling advanced functionality such as **competence scoring**, **question generation**, **translation feedback**, and **instructional adaptation**. It plays a central role in assessing user input, generating dynamic content, and personalizing the learning experience for German language learners.

The LLM is used in several key interaction contexts:

- It evaluates **speaking and writing competence** by comparing user input, whether spoken (ASR transcript) or written (typed) against a ground truth reference. The system then assigns a score between 0 and 10 based on correctness, fluency, and task completion.
- It supports **translation training** by analyzing user-provided German or English equivalents of a source text and offering corrections or stylistic improvements. This allows learners to practice both L1→L2 and L2→L1 translation while receiving feedback beyond simple literal matching.
- For **reading and listening comprehension**, the LLM generates **contextual questions** based on a short German paragraph (either written or spoken by TTS). Learners must answer these questions to demonstrate understanding, and the LLM evaluates the quality and relevance of their responses.
- It provides **instructional scaffolding** by rephrasing content, giving hints, or simplifying/recomplicating tasks depending on the learner's current ability.

Technically, the system can run with any sufficiently advanced LLM. The know-how is in the prompts and the chaining of LLM agents (Chu et al. 2025) to achieve better outcomes as with single-shot prompts. Since multilingual training of LLMs is the standard in the last year and not only US-developed models like Llama are proficient in German language (Peinl & Wirth 2024), but German and other European languages are also standard for Chinese LLMs like Qwen and DeepSeek, there is abundant choice (Phogat et al. 2025). We have used local installations of Llama 3.3 70B and Mistral 3.2 small up to now, besides tests with GPT-4o and Claude 3.7 Sonnet. For the future, an upgrade of our local LLM infrastructure to GLM 4.5 Air or Qwen 3 235B A22 2507 is planned. We did tests with a locally installed Phi-4 reasoning as well but did not get better results compared to Llama.

4. Use Cases

4.1 Speaking Tasks

The system presents the learner with a short text to read out aloud. This text is adjusted to the learning level of the learner (A1..B2) and previous mistakes, so that problematic words occur more frequently than others. The recording is then transcribed using both ASR systems and deviations from the expected result are displayed to the user. **Figure 3** shows example feedback. Please note that the system concentrates on words not understood by G-ASR and ignores the difference in pronunciation for "Waschmaschine". If the WER for G-ASR is zero, the phoneme-level mistakes are also highlighted, except for phonemes from a similarity list.

Ground Truth Grapheme:	meine hose ist schmutzig hast du eine waschmaschine
Recognized Grapheme:	meine hause ist schmierstich hast du eine waschmaschine
Grapheme WER: 25%	Grapheme CER: 17.6%
Ground Truth Phoneme:	ˈmaɪ̯nə ˈhoːzə ɪst ˈʃmʊ͡tsɪk. hast duː ˈaɪ̯nə ˈvaʃmaʃiːnə?
Recognized Phoneme:	ˈmaɪ̯nə ˈhaʊ̯zɐ ɪst ˈʃmjuːʃtɪç hast duː ˈaɪ̯nə ˈvɔʃemaʃiːnə?
Phoneme WER: 37.5%	Phoneme CER: 23.7%

Figure 2: Example of feedback on speaking tasks (all words are lower cased for comparison).

4.2 Listening Tasks

The system presents the learner a short audio file that was produced with our in-house TTS system. The audio files are prerecorded and not generated on demand, so that there is a chance for manual quality assurance. The task can be either to transcribe the audio word by word, or to write down the semantic meaning of the spoken information. In the first case, spelling is scored together with listening comprehension and the system cannot distinguish which kind of mistake was made. Feedback is given regarding the character error rate (CER).

Table 1: Example of feedback on listening and spelling task.

Groundtruth	User input	CER	Time [s]
Ich habe die Milch verschüttet.	Ich habe die mich verschuldet.	13.3%	55.7
400 Euro Miete im Monat sind viel Geld.	400 Euro mitte im Monat sind vier Geld.	4.3%	30.1
Weihnachten ist in Deutschland ein besonderer Feiertag.	weihnachten ist deutschland ein besonderer freitag	11.1%	21.3
Kannst Du mir eine Geschichte vorlesen?	Kannt su mir eine Geschichte vorlesen?	5.3%	23.2
Meine Wohnung hat zwei Zimmer, Küche und Bad.	Am wohnung hatz wer markische im park	48.8%	59.3

In the second case, an LLM is used to score the semantic similarity between the learners' input and the spoken text. It can be prompted to accept input in German only, or to accept single words or whole sentences in different languages. In our primary use case, incoming students from India are e.g. proficient English speakers and therefore we accept English input for the listen and understand task (see **Figure 4**).

Table 2: Example of feedback on listening comprehension tasks (scoring by Llama 3.3 70B).

Groundtruth	User input	Rating [0..10]
400 Euro Miete im Monat sind viel Geld.	Vierhundert Euro Miete pro Monat ist teuer!	9
	400€ Miete im Monat sind viel Geld.	10
	viele hundert euro mieten im Monat viel Geld	6
	Four hundred euros a month for rent is a lot of money.	10
Weihnachten ist in Deutschland ein besonderer Feiertag.	Weihnachten ist ein wichtiger Feiertag in Deutschland.	9
	Weihnachten ist für Deutschland very special.	8
	Weihnachten is in deutsche land bisonde feir	2
	Christmas is a special feast day in Germany	9

4.3 Writing Tasks

In written form, a lot of different training types are possible and many of them are already implemented in our system. The main focus is on creating the training tasks automatically by an LLM instead of manually by a human expert. We use text from websites that provide contents for German children or specifically for German language learners as a basis (e.g. helles-koepfchen.de, or “stories in easy language” by PH Ludwigsburg). We break the text down into single sentences and further filter them according to a list of words that should be known at a certain language level. Based on these texts, the LLM generates different training tasks in a multi-stage pipeline.

For “**fill-in-the-blanks**” kind of tasks, the challenge is to find meaningful words as alternatives for the correct word. These alternatives should not be obviously wrong even for people with low language competency level but should be not too similar so that the correct solution can no longer be determined unambiguously. Tests with Llama 3.3 70B, GPT-4o, and Claude 3.7 Sonnet showed that one-shot solutions were inadequate. Therefore, we developed a pipeline that first generates candidate words and then tests each single one as part of the original sentence and the LLM decides whether it is a valid alternative according to the criteria or not. We repeat up to three times, skipping sentences without sufficient alternatives. We allow up to two blanks per sentence, if the sentence has at least eight words. Otherwise, only a single blank is inserted. Task difficulty can be varied by providing multiple sentences with blanks at once that share a common pool of word candidates.

Grammar training is similar to fill-in-the-blank, but learners receive only the base form and must supply the correct inflection. This is suitable for verbs as well as nouns. To make generation easier, only the basic form is provided in our tasks with no explicit choices.

The insertion of definite articles (“der”, “die”, “das” in German in contrast to “the” in all cases in English) is another slight variation of the task that addresses a common problem for German language learners, since the grammatical gender of a word is often counter-intuitive. It is e.g. “das Mädchen” (the girl) with a neutral article although you would expect a female gender for girl. The other way around, milk has a female gender in German although most other drinkable things are neutral.

Translation tasks are easier to produce. The capabilities of general LLMs in machine translation of high-resource languages is already very good (Kocmi et al. 2024) and the German – English language pair is one of the best (Manakhimova, et al. 2024), only topped by French – English and Spanish – English. As known from other approaches like LLM-as-a-judge (Zheng et al. 2023), the LLMs are even better at judging a given text regarding its correctness than producing a correct text for a given task on their own. This can also be transferred to the machine translation domain. (Treviso et al. 2024) show that LLMs can give qualified feedback to translations. We employ the same principle and let the LLM judge the results from the language learners output for both English to German as well as German to English translation tasks. Especially for the simple sentences of A1 and A2 level, the error rate of this approach is extremely low and can be neglected.

Question answering is another way of practicing a new language and demonstrate language understanding. Since this task is also important for the development of LLMs, it is well studied there for both answering questions as well as generating meaningful questions (Li & Zhang 2024). Although some challenges remain (Al Faraby & Romadhony 2024), a multi-stage pipeline of generation, review and selection of questions for a given text leads to good results and minimizes chances of questions that are not answerable by reading the text alone

or have ambiguous answers. (Biancini, Ferrato, Limongelli 2024) find that even GPT-3.5 can generate meaningful multiple-choice questions within an educational context. Our tests show that both Llama 3.3 70B as well as Mistral 3.2 small are able to generate good questions in a single step, if the prompt is sophisticated enough.

5. Discussion

A preliminary evaluation using only G-ASR was conducted with 24 L2 speakers to assess the feasibility of automated pronunciation assessment. The results highlight the inherent subjectivity in pronunciation judgment, which complicates the creation of a reliable ground truth. The pair-wise inter-rater agreement between three human raters (all German native-speakers) was between 0.71 (rather good) and 0.46 (rather bad). We tested the correlation between both single human raters and several different G-ASR systems and found a significant negative correlation between the average WER or three ASR systems and the average of the human raters (-0.593), with a medium to strong confidence. These results suggest that lower WERs are generally associated with higher human-assessed pronunciation quality, an expected but useful validation of G-ASR-based scoring.

Building on this initial insight, we developed and piloted a more specialized system incorporating P-ASR to address the limitations of G-ASR in capturing detailed articulatory patterns. Although the system has so far only been tested in a small-scale pilot study involving 10 incoming students from India, it has already demonstrated promising usability for assessing competencies in listening and speaking. Each participant completed 15 speaking tasks (see Section 4.1) and 15 listening tasks (see Section 4.2). Their responses were evaluated both automatically by the system (with P-ASR and G-ASR combined) as well as manually by two native-speaking evaluators. While minor discrepancies between automatic and manual scoring remained, incorporating the P-ASR module resulted in a more accurate reflection of learners' speaking abilities compared to the use of G-ASR alone. In some instances, however, P-ASR overcorrected by transcribing mispronounced words as their intended phoneme representations, which suggests possible overfitting to the training data. To address this, earlier model checkpoints should be reviewed to determine whether they produce phoneme transcriptions that more accurately reflect the phoneme representation of the actual speech input. A more extensive evaluation with a larger sample and broader assessment criteria is needed to quantitatively assess the system's effectiveness in productive use. Additionally, while the ASR systems used are state-of-the-art (SOTA), a small risk remains that errors in recognition do not stem from bad pronunciation or other mistakes of the language learner at all, but from errors of the ASR model. However, in quiet environments and under good recording conditions, this error will be minimal, which can be seen from the low WER of LibriSpeech clean (1.5%) and the HUI dataset for German language (1.9%, Wirth & Peinl 2022).

6. Conclusion

In this work, we present TalkPro, a modular, AI-based language learning system designed to support and assess German L2 learners. The system integrates phoneme- and grapheme-level ASR, multi-speaker TTS, and LLM-based components, combining these modules in multiple ways to support training and assessment across key language competencies, including listening, pronunciation, translation, and comprehension.

While the current version already supports most previously described features, it has so far only undergone a limited pilot phase with a small number of users. As part of future work, we plan to conduct a larger-scale evaluation involving incoming Indian students enrolled in master's programs in Germany. This will allow us to assess both the overall effectiveness of the system and the specific contribution of individual modules and exercises. We aim to measure learning outcomes more precisely, explore usage patterns in authentic educational settings, and refine the feedback mechanisms based on empirical data. As part of this refinement, we will also investigate the performance of earlier P-ASR model checkpoints to identify a version that balances phoneme-level sensitivity with tolerance for learner variation, thereby avoiding overcorrection and better representing the actual pronunciation of speech input.

Preliminary experiments with incoming students from India to Germany indicate that phoneme-level ASR effectively identifies pronunciation errors, whereas grapheme-level ASR tends to overlook them.

Future research will involve a comprehensive evaluation of automated results against human judgment, alongside the expansion of TalkPro's training capabilities with LLM-based reading comprehension modules that prioritize conceptual understanding over traditional verbatim recall.

Ethics Declaration

Ethical clearance was not required for the research presented in this work.

AI Declaration

The authors have used generative AI solely for the purpose of rephrasing in a few cases in section 2 and take full responsibility for the content.

References

Al Faraby, S., & Romadhony, A. (2024). Analysis of LLMs for educational question classification and generation. *Computers and Education: Artificial Intelligence*, *7*, 100298.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference (pp. 4218-4222)*.

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems. *Innovation in Language Learning and Teaching*, 18(5), pp. 443–461. Available at: https://doi.org/10.1080/17501229.2024.2315101.

Biancini, G., Ferrato, A., & Limongelli, C. (2024). Multiple-choice question generation using large language models: Methodology and educator insights. *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 584-590).

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., ... & Wen, Q. (2025). LLM Agents for Education: Advances and Applications. arXiv. Available at: https://doi.org/10.48550/arXiv.2503.11733.

Hişmanoğlu, M. (2006) Current perspectives on pronunciation learning and teaching, *Journal of language and linguistic studies*, 2(1), pp. 101–110.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Interspeech 2020*.

Honnibal, M. *et al.* (2020) spaCy: Industrial-strength Natural Language Processing in Python. Available at: https://doi.org/10.5281/zenodo.1212303.

Kim, J., Kong, J. & Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, in M. Meila and T. Zhang (eds) *Proceedings of the 38th International Conference on Machine Learning*. PMLR (Proceedings of Machine Learning Research), pp. 5530–5540.

Kim, S., Le, D., Zheng, W., Singh, T., Arora, A., Zhai, X., ... & Seltzer, M. L. (2021). Evaluating user perception of speech recognition system quality with semantic distance metric. *arXiv preprint arXiv:2110.05376*.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., ... & Shmatova, M. (2024). Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation* (pp. 1-46).

Lee, J., Jang, J. and Plonsky, L. (2015). The Effectiveness of Second Language Pronunciation Instruction: A Meta-Analysis. *Applied Linguistics*, 36(3), pp. 345–366. Available at: https://doi.org/10.1093/applin/amu040.

Li, K. & Zhang, Y. (2024). Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation. in L.-W. Ku, A. Martins, and V. Srikumar (eds) *Findings of the Association for Computational LinguisticsFindings 2024*, Bangkok, Thailand: Association for Computational Linguistics, pp. 4715–4729.

Liu, Y., binti Ab Rahman, F., & binti Mohamad Zain, F. (2025). A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Education*, *12*(1), 2466288.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293-311.

Manakhimova, S., Macketanz, V., Avramidis, E., Lapshinova-Koltunski, E., Bagdasarov, S., & Möller, S. (2024). Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation* (pp. 355-371).

Moxon, S. (2024). All-talk: Enhancing EFL pronunciation with Microsoft azure speech services. *Abac Journal*, *44*(4), 139.

Müller, T. & Kreutz, D. (2022). ThorstenVoice Dataset 2022.10. Available at: https://doi.org/10.5281/zenodo.7265581.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*(5), 393-408.

Neri, A., Cucchiarini, C., & Strik, H. (2002). Feedback in computer assisted pronunciation training: When technology meets pedagogy. Available at: https://www.academia.edu/download/40482205/neri.2002.2.pdf

Pennington, M.C. (2021). Teaching Pronunciation: The State of the Art 2021. *RELC Journal*, 52(1), pp. 3–21.

Pennington, M.C. & Rogerson-Revell, P. (2018). English Pronunciation Teaching and Research: Contemporary Perspectives. Palgrave Macmillan.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*.

Puchtler, P., Wirth, J. & Peinl, R. (2021). HUI-Audio-Corpus-German: A high quality TTS dataset. *44th German Conference on Artificial Intelligence (KI2021)*. Berlin, Germany.

Puvvada, K. C., Żelasko, P., Huang, H., Hrinchuk, O., Koluguri, N. R., Dhawan, K., ... & Ginsburg, B. (2024). Less is More: Accurate Speech Recognition & Translation without Web-Scale Data. *Interspeech 2024* (pp. 3964-3968).

Şentürk, R. (2023). Die Rolle Künstlicher Intelligenz beim Deutsch als Fremdsprachenlernen: Eine Untersuchung am Beispiel von ChatGPT. Diyalog Interkulturelle Zeitschrift für Germanistik, 11(2), pp. 405–430.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... & Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *11th Intl Joint Conf. on Natural Language Processing*. Association for Computational Linguistics, pp. 993–1003.

Wirth, J. (2025). OLaPh: Optimal Language Phonemizer. Available at: https://doi.org/10.48550/arXiv.2509.20086.

Wirth, J. & Peinl, R. (2022). Automatic Speech Recognition in German: A Detailed Error Analysis. *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. pp. 1–8