Read to Hear: a Zero-Shot Pronunciation Assessment using Textual Descriptions and LLMs

Anonymous ACL submission

Abstract

Automatic pronunciation assessment is typically performed by acoustic models trained on audio-score pairs. Although effective, these systems provide only numerical scores, without the information needed to help learners understand their errors. Meanwhile, large language models (LLMs) have proven effective in supporting language learning, but their potential for assessing pronunciation remains unexplored. In this work, we introduce TextPA, a zero-shot, Textual description-based Pronunciation Assessment approach. TextPA utilizes human-readable representations of speech signals, which are fed into an LLM to assess pronunciation accuracy and fluency, while also providing reasoning behind the assigned scores. Finally, a phoneme sequence match scoring method is used to refine the accuracy scores. Our work highlights a previously overlooked direction for pronunciation assessment. Instead of relying on supervised training with audio-score examples, we exploit the rich pronunciation knowledge embedded in written text. Experimental results show that our approach is both cost-efficient and competitive in performance. Furthermore, TextPA significantly improves the performance of conventional audio-score-trained models on out-of-domain data by offering a complementary perspective.

1 Introduction

Automatic pronunciation assessment offers an alternative to traditional language instruction by providing learners with accessible, scalable, and timely feedback on their speaking abilities. Most prior work in this area relies on supervised learning: collecting speech recordings annotated with pronunciation scores from human instructors and training acoustic models to assess proficiency scores (Chen et al., 2024; Gong et al., 2022). Although effective, models trained on audio-score pairs provide only numerical scores, offering little insight into why a particular score was assigned. Collecting more informative and descriptive feedback, such as detailed comments from human raters, can be time-consuming and expensive.

Recently, Large Language Models (LLMs) have gained popularity for their ability to generate natural, context-aware responses. We propose that this generative capability can be leveraged to produce explainable feedback in pronunciation assessment, going beyond simple scoring. LLMs have also demonstrated the potential to provide valuable insights into language learning (C Meniado, 2023). Most studies focus on using LLMs in writing tasks (Lo et al., 2024). However, LLMs also capture knowledge of language speaking, as humans have documented their knowledge about pronunciation in written form to facilitate sharing and teaching. Also, previous studies have shown that LLMs, such as GPT, have the potential to interpret textual descriptions of speech signals. In (Wang et al., 2023), researchers wrote the pause durations in a sentence – e.g., "it (<10 ms) is (<10 ms) nothing (10 ms–50 ms) like (<10 ms) this," – and put the sentence into GPT to assess whether the pauses are correct. However, this study focused only on detecting inappropriate pauses using duration information, without exploring the ability of LLMs to interpret other key dimensions of pronunciation, such as articulation or intonation.

To bridge the gap between the textual understanding of LLMs and the physical acoustic signal, audio-language models (ALMs) (Elizalde et al., 2023; Tang et al., 20234; Chu et al., 2023) have emerged. ALMs integrate audio and text by encoding audio into audio tokens, which are then processed by the LLM with text tokens. However, most open-source ALMs are pre-trained on audio captioning or speech recognition datasets and show limited ability to assess speech without fine-tuning (Deshmukh et al., 2024; Wang et al., 2025b). Also, due to computational constraints, these studies use smaller LLMs (e.g., 7B or 13B Llama), limiting their ability to fully leverage LLM capabilities. Closed-source large ALMs such as GPT-audio and Gemini-audio have demonstrated the potential for pronunciation assessment in zero-shot settings (Wang et al., 2025a), but large ALMs are costly to operate with an audio input. Since audio tokens are much more expensive than text tokens[1] and the number of audio tokens generated from a speech signal can be much greater than the number of text tokens in its corresponding transcript, using a large ALM with audio inputs is considerably more expensive than using LLM with text inputs.

We explore an alternative method to bridge the gap between LLM's textual knowledge and physical speech signals. Instead of relying on audio tokens, our method uses the existing capabilities of LLMs by selecting text-based acoustic descriptors common in written text. Pre-trained acoustic models are used to generate these, including transcripts, phoneme sequences (in both International Phonetic Alphabet (IPA) and CMU Pronouncing Dictionary (CMU) formats), and pause durations. The descriptors are provided as input to LLMs for pronunciation assessment. We incorporate a similarity score between the recognized IPA sequence and the transcript-mapped canonical IPA sequence to improve assessment of pronunciation accuracy.

Our contributions are: 1) We propose TextPA, a zero-shot pronunciation assessment model that uses textual descriptions of speech signals; 2) Our method produces interpretable and explainable feedback, unlike conventional pronunciation assessment systems that yield only numeric scores and incorporating TextPA enhances the performance of an audio-score-trained model on out-of-domain data; 3) Compared to large ALMs, our approach significantly reduces API costs while delivering competitive or superior assessment performance.

2 TextPA

To assess English pronunciation in terms of accuracy and fluency. Textual acoustic cues are extracted using a set of pre-trained models: the transcript is obtained from an automatic speech recognition (ASR) model; pause information and the recognized CMU sequence are derived from a phonetic aligner; and the IPA phoneme sequence is generated using a phoneme recognition model. These textual representations are then provided as input to an LLM, which is prompted to assess the pronunciation and produce both accuracy and fluency scores, along with the reasoning behind its evaluations. Lastly, we introduce IPA match scoring to further refine the accuracy score. Figure 1 presents an overview of this TextPA framework. Our approach operates in a zero-shot setting, relying only on existing pre-trained acoustic models and LLMs, and does not require any audio-score pronunciation data.

2.1 Textual Acoustic Cues for LLM Input

2.1.1 Transcript

A transcript lacking semantic coherence may result from inaccurate recognition due to poor pronunciation. Repeated words within a sequence or filler words such as *"hmm,"* can indicate a lack of fluency. In Figure 2, the speaker is told to say *"his head hurts even worse,"* but their pronunciation is highly inaccurate. Except for *"His."*, all other words received only *3* out of *10* points. Due to poor pronunciation and lack of fluency, the ASR model produced an inaccurate transcript (i.e., *"His hand hands very well"*) which is semantically incoherent, signaling low pronunciation proficiency for the LLM, as reflected in its reasoning. However, as the ASR model is designed to recognize words rather than analyze pronunciation, it may automatically correct inaccurately pronounced words to produce a semantically coherent sentence. For example, in Figure 3, the speaker is instructed to say *"maybe we should get some cake"* but mispronounced *"cake."* Although the pronunciation is inaccurate, the ASR transcript (*"maybe we should get some cards,"*) is still semantically reasonable. As a result, using the transcript alone provides limited insight into the finer details of articulation. To address this, we incorporate the IPA and CMU phoneme sequences that explicitly represent spoken sounds.

2.1.2 Recognized IPA and CMU Phoneme Sequence

IPA, widely used in linguistics, dictionaries, and language education materials, is a standardized phonetic notation system that represents the sounds of spoken language using a consistent set of symbols. Each symbol corresponds to a specific speech sound, providing a one-to-one mapping between sound and notation. The CMU phoneme sequence is a phonetic transcription format based on the

[1]For example, the OpenAI *GPT-4o-mini-audio* model charges $10.00 per 1M audio tokens, compared to $0.15 per 1M text tokens (as of April 2025).

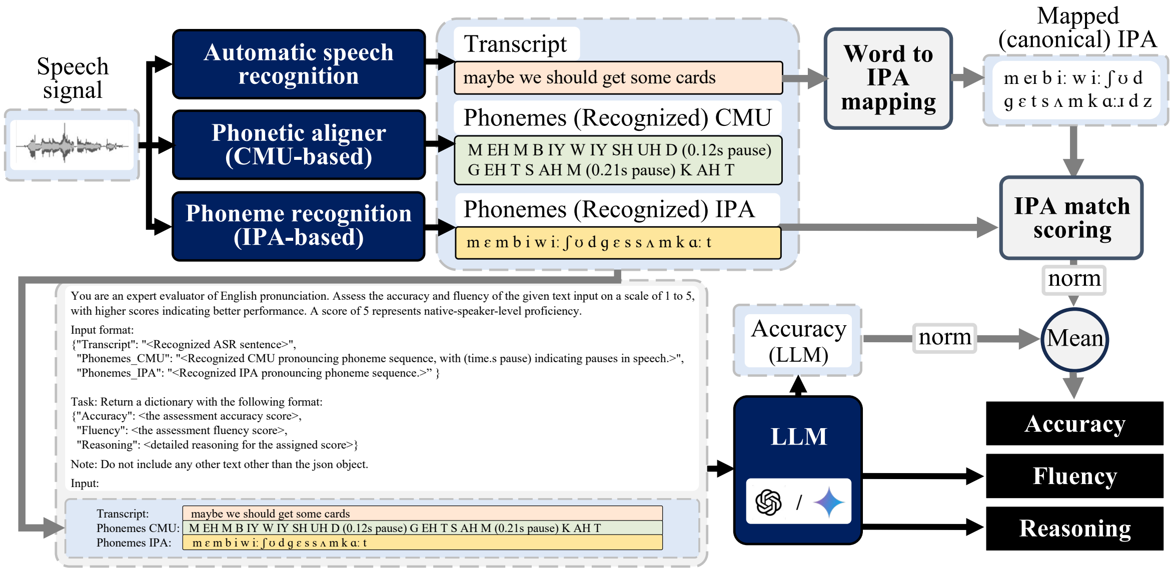

Figure 1: An overview of TextPA.

Carnegie Mellon University Pronouncing Dictionary (CMUdict). Unlike IPA, which is universal in language and more fine-grained, CMU uses a simplified set of phonemes tailored for American English, which is widely used in speech processing applications due to its compatibility with ASR systems and phoneme-based models. Because both representations are widely used, LLMs trained on extensive text corpora have encountered and internalized the mapping between IPA and CMU phoneme annotations and the word. For example, in Case Study B (Figure 3), by comparing the recognized IPA and CMU sequences, the LLM can identify that the word "cards" may have been mispronounced and can leverage this information to assess pronunciation accuracy. It is able to align transcript words with the corresponding phoneme sequence, even when word boundaries are not explicitly marked. We also embed pause information from the phonetic aligner into the recognized CMU phoneme sequence. Pauses are annotated in an easily interpretable format, e.g. *"D (0.12s pause) G"* indicates a *0.12-second pause* between the phones *"D"* and *"G"*. As shown in Case Study B (Figure 3), the LLM leverages this pause information when reasoning about the speaker's fluency.

2.2 IPA Match Scoring

To assess pronunciation, the LLM internally maps each word in the transcript to its canonical phoneme sequence and compares it with the provided recognized phoneme sequence. Although LLMs are capable of this, as shown in Case Study B (Figure 3) where the model correctly identifies the mispronunciation of the word *"cards"*, they may still overlook some errors. For example, in the same case, a discrepancy is observed between the canonical phoneme sequence for the word *"maybe"* (m eɪ b iː / M EY B IY) and the recognized sequence (m ɛ m b i / M EH M B IY), indicating inaccurate pronunciation. Although the human annotation assigns a score of 10 out of 10 to the pronunciation of *"maybe"*, our manual inspection suggests that the word is not clearly articulated. However, the LLM does not reflect this error in its reasoning.

To further refine accuracy assessment, we introduce IPA match scoring, which measures the similarity between the recognized and canonical phoneme sequences and uses this as an indicator of pronunciation accuracy.[2] To perform IPA match scoring, the transcript is first mapped to its canonical IPA phoneme sequence using a pronunciation dictionary. We then compute the similarity between the transcript-mapped and the recognized IPA sequences using the Smith-Waterman algorithm (Smith et al., 1981). This algorithm is a dynamic programming method originally developed for local alignment of biological sequences such as DNA or proteins. It identifies the most similar subsequences between two inputs, making

[2]We use IPA instead of CMU because IPA offers more fine-grained phonetic detail. In addition, our empirical results suggest that match scoring using IPA consistently outperforms scoring with CMU phonemes.

Case study A

Ground-truth (Human annotation)	his head hurts even worse (10, 3, 3, 3, 3,)
LLM input	
Transcript	his hand hands very well
Phonemes CMU	HH IH Z HH AE N D HH AE N D Z (0.05s pause) ER N EH R IH IY R IH IY AH L
Phonemes IPA	h ɪ z h æ n d h æ n d s m ɛ ɹ i ɹ ɪ
Mapped IPA	h ɪ z h æ n d h æ n d z v ɛ ɹ i w ɛ l

LLM reasoning

The accuracy is low due to several mispronunciations and inaccuracies in the phoneme recognition. **"very well" is completely different from "ER N EH R IH IY R IH IY AH L" and "m ɛ ɹ i ɹ ɪ".** The word "hands" is repeated and there's a pause. The fluency is slightly affected by the pause and **the unnatural sequence of words.**

Figure 2: Case study A. Due to the inaccurate pronunciation, the ASR model produced unnatural word sequences, which in turn signaled to the LLM that the pronunciation was not fluent.

Case study B

Ground-truth (Human annotation)	maybe we should get some cake (~~10~~, 10, 10, 10, 10, 3,) *incorrect
LLM input	
Transcript	maybe we should get some cards
Phonemes CMU	M EH M B IY W IY SH UH D (0.12s pause) G EH T S AH M (0.21s pause) K AH T
Phonemes IPA	m **ɛ** **m** b **i** w iː ʃ ʊ d g ɛ **s** s ʌ m k **ɑː t**
Mapped IPA	m **eɪ** b **iː** w iː ʃ ʊ d g ɛ **t** s ʌ m k **ɑːɹ d z**

LLM reasoning

Accuracy: The pronunciation is generally accurate, with most phonemes being correctly produced. **The vowel in 'cards' is slightly off. It appears to be truncated to 'cot' sound rather than 'cards'.** Still quite understandable.
Fluency: **The presence of pauses affects the fluency score. (0.12s) is acceptable, (0.21s) pause in a short sentence is a bit long.** The pace is slightly uneven overall. More natural speech would exhibit smoother transitions between words.

Figure 3: Case study B. The transcript is semantically reasonable despite inaccurate pronunciation. To address this, the mismatch between the recognized phoneme sequence and the transcript provides the LLM with insight into potential articulation inaccuracies. The mapped IPA (i.e., the canonical IPA of the transcript) is shown for reference and is not provided as input to the LLM.

it well-suited for phoneme sequence comparison, which consists of subsequences representing different words. Since the IPA-matched scores and the LLM-generated accuracy scores operate on different scales, we apply min-max normalization across the test set to align them within a common range. The final accuracy score is then calculated as the average of the normalized Smith-Waterman similarity and the normalized LLM-generated score.

3 Experimental Setup

3.1 Data and Evaluation Metric

We evaluated TextPA on the open-source Speechocean762 (Zhang et al., 2021) and MultiPA (Chen et al., 2024) datasets [3], both of which focus on English speech produced by native Mandarin speakers. The Speechocean762 (abbreviated as Speechocean) dataset consists of 5,000 utterances spoken by 250 speakers, with annotations at the sentence, word, and phoneme levels. In this study, we focus on sentence-level accuracy, fluency, and prosody. The utterances in Speechocean are scripted. Participants were instructed to read predefined sentences, making the ground-truth transcript available. However, our method operates without the need for ground-truth information. Most sentences in Speechocean are short, as shown in Figure 1, 2, and 3, with corresponding audio durations ranging from 2 to 20 seconds. Since TextPA requires no training, we used only the Speechocean test set, which contains 2,500 utterances.

The MultiPA data contains 50 audio clips, each ranging from 10 to 20 seconds in duration, collected from ~20 anonymous users interacting with a dialogue-based chatbot. Unlike Speechocean, where speakers are asked to read predefined sentences, MultiPA data captures open-ended responses, allowing learners to speak freely or answer questions. This allows for a more authentic assessment of learners' speaking abilities. Table 1 shows example transcriptions from both datasets. We use the Pearson correlation coefficient (PCC) as the main evaluation metric since it has often been used in prior studies and provides better interpretability

[3]License: Attribution 4.0 International (CC BY 4.0)

Speechocean	Two, four, seven.
	It was good for me.
MultiPA data	I'm an active person and I enjoy playing a variety of sports. One of my favorite sports to play is basketball as it is a great way to stay fit and socialize with friends at the same time.
	I often go to the zoo. I think the zoo is a very interesting place. And I go, I went to the zoo once a week now.

Table 1: Example transcriptions from Speechocean and MultiPA. Speechocean consists of relatively short, scripted utterances from read-aloud tasks, whereas MultiPA data captures open-ended, conversational speech.

when comparing performances on different dataset.

3.2 Implementation Details

We use Whisper (Radford et al., 2023) (*large-v3-en*) for transcription, the model from (Xu et al., 2021)[4] for IPA sequence, Charsiu (Zhu et al., 2022) predictive aligner for CMU sequence, and Phonemize (Bernard and Titeux, 2021)[5] for word-to-IPA mapping. Acoustic models were run on an NVIDIA RTX 4500 GPU. The LLM uses default API settings, and results are from a single run.

4 Results

4.1 Performance on Free-speech

Table 2 shows the performance on MultiPA data. We compare TextPA with different LLM back-ends. Because TextPA *(gpt-4o-mini)* performs better than the TextPA *(gemini-2.0-flash)*, we chose to run GPT-4o-mini-audio to compare the performance. Results suggest that the proposed TextPA outperforms GPT-4o-mini-audio in assessing pronunciation, achieving better performance in both accuracy and fluency. We also compare performance with the MultiPA model (Chen et al., 2024), an acoustic model trained on Speechocean. Results show that the proposed TextPA achieves higher accuracy and provides competitive fluency assessment, showing the effectiveness of TextPA in a zero-shot setting.

We evaluate the effectiveness of combining the MultiPA and TextPA models. To account for differences in the scale of their prediction scores, we first apply min-max normalization to each model's outputs. The final prediction is obtained by averaging the normalized scores. Despite the simplicity of this fusion strategy, the combined model achieves notable performance improvement over using either model alone. This improvement is likely due to the distinct sources of information. MultiPA is trained on paired audio-score data, learning directly from acoustic examples, whereas TextPA operates solely on text and leverages prior knowledge about pronunciation assessment. Differing approaches offer diverse perspectives, enabling the combined system to achieve improved performance.

Due to the limited amount of paired audio-score pronunciation data, MultiPA may have difficulty accurately assessing words that were not encountered during training. In contrast, TextPA has access to a much broader vocabulary, leading to higher performance on accuracy assessment. However, because MultiPA analyzes raw audio recordings, it can capture acoustic cues such as detailed phone-level durations or pitch variations. These cues are typically not represented in written descriptions or are difficult to capture accurately in text, making them challenging for LLMs to interpret. In fact, we also explore the LLM's ability to assess prosody using ToBI annotations (Beckman and Hirschberg, 1994) which offer a text-based representation of tonal patterns and phrase boundaries. However, the LLM appears to struggle with assessing prosody by accurately interpreting these annotations, even when given explicit instructions (see the Appendix B for details). In essence, the two approaches provide complementary advantages on the assessment task, and combining them could be beneficial by leveraging the strengths of both.

	Accuracy	Fluency
TextPA *(gemini-2.0-flash)*	0.697	0.557
TextPA *(gpt-4o-mini)*	**0.728**	0.650
GPT-4o-mini-audio	0.674	0.648
MultiPA model	0.618	**0.683**
MultiPA model + TextPA *(gpt-4o-mini)*	**0.769**	**0.784**

Table 2: Model performance on MultiPA data. Note that MultiPA model was trained on Speechocean.

4.2 Performance on Scripted Utterances

Table 3 shows the performance on Speechocean. We first compare the performance of TextPA

[4] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

[5] *EspeakBackend("en-us")*

using different LLM back-ends. Results indicate that *gemini-2.0-flash* outperforms *gpt-4o-mini*; therefore, we conduct another experiment for the Gemini-2.0-flash-audio in our performance comparison. In contrast to its strong performance on the MultiPA dataset, TextPA performs relatively poorly on Speechocean. This discrepancy might arise from fundamental differences between the datasets. Speechocean consists of shorter, more constrained utterances (as shown in Table 1) which offer limited phonetic and semantic variation. Moreover, Speechocean prompts students to repeat predefined sentences, unlike the MultiPA data, which produces free-form speech. As a result, both the pause cues between words and the semantic content of the transcripts offer weaker indicators of language proficiency, thereby reducing the effectiveness of TextPA. These dataset differences may also explain the performance inconsistency between Gemini and GPT across the two datasets. Nevertheless, TextPA remains competitive on Speechocean. Note that TextPA relies solely on text tokens, whereas Gemini-2.0-flash-audio uses text tokens for instructions and audio tokens for input speech signals. [6] We also include in-domain models as references. Since TextPA is a zero-shot approach without using training data, the in-domain models naturally perform better. Directly combining the predictions as done with MultiPA data does not lead to improvements for the in-domain setting due to the performance gap. Further investigation is needed to explore more effective ways of leveraging TextPA for in-domain models.

	Accuracy	Fluency
Zero-shot		
TextPA *(gpt-4o-mini)*	0.507	0.466
TextPA *(gemini-2.0-flash)*	0.532	**0.557**
Gemini-2.0-flash-audio	**0.562**	0.556
In-domain		
(Lin and Wang, 2022)	0.72	-
(Liu et al., 2023b)	-	0.795
MultiPA model	0.705	0.772

Table 3: Model performance on Speechocean.

4.3 Ablation Study on Textual Descriptions of Speech Signals

We evaluate the performance of accuracy scoring based on phoneme sequence matching. Our findings demonstrate that IPA match scoring is a straightforward yet highly effective method for assessing pronunciation accuracy. We also investigated the performance of CMU match scoring. Similar to IPA match scoring, the words in the transcript are mapped to CMU labels using the dictionary, and then compared with the recognized CMU sequence through normalized Smith-Waterman similarity scores. However, the results indicate that the CMU sequence is less effective for accuracy assessment compared to the IPA sequence. This difference comes from the higher level of detail in the pronunciation representation of the IPA, which contains more than 107 syllable letters, while the CMU set contains only 39 phonemes.

We also performed an ablation study to determine which textual descriptions of acoustic cues are most effective for language models in pronunciation assessment. When using an LLM, the transcript alone can offer insights. Augmenting the input with recognized IPA sequences improves performance, particularly in accuracy, as the LLM can compare word transcriptions with their phonetic transcriptions to better identify mispronunciations. Adding CMU sequences alongside the transcript helps to enhance both accuracy and fluency as well: accuracy improves for similar reasons as with IPA, while fluency benefits from the pause information encoded in CMU sequences. Overall, combining the transcript, CMU, and IPA sequences leads to the best performance, with IPA match scoring providing additional boosts in accuracy.

4.4 Impact of ASR Transcription Quality

Transcripts play a crucial role in TextPA. To examine the affect of ASR model quality (i.e., transcript quality), we compared LLM-based assessment using transcripts generated by two Whisper variants: *large-v3-en* (denoted as *large-en*) and *tiny*. The *large-en* model, with 1550M parameters, is English-only and generates higher-quality transcripts that are more robust to inaccurate pronunciation. In contrast, the *tiny* model, with only 39M parameters and multilingual training, is more likely to produce transcription errors or misclassify

[6]The cost of *gemini-2.0-flash* is 0.1 per 1M text tokens and $0.7 per 1M audio tokens, making Gemini-2.0-flash-audio approximately 3.5 times more expensive in API calls than running TextPA *(Gemini-2.0-flash)* on the Speechocean.

MultiPA data		
	Accuracy	Fluency
TextPA *(gpt-4o-mini)*	0.728	0.650
LLM: all	0.643	0.650
LLM: trans.+cmu	0.491	0.485
LLM: trans.+ipa	0.452	0.410
LLM: transcript	0.404	0.432
IPA match scoring	0.653	-
CMU match scoring	0.208	-
Speechocean		
	Accuracy	Fluency
TextPA *(gemini-2.0-flash)*	0.532	0.557
LLM: all	0.456	0.557
LLM: trans.+cmu	0.427	0.553
LLM: trans.+ipa	0.448	0.458
LLM: transcript	0.313	0.310
IPA match scoring	0.507	-
CMU match scoring	0.263	-

Table 4: Ablation study of text-based acoustic cues. We selected the LLM with the best performance on each dataset as the representative model: *gpt-4o-mini* for the MultiPA data and *gemini-2.0-flash* for the Speechocean data. *LLM: transcript* uses only the transcript as input. *LLM: trans.+ ipa* and *trans.+ cmu* add IPA or CMU sequences, respectively. *LLM: all* combines all three inputs: transcript, IPA, and CMU. Note that the fluency scores for *LLM: all* and TextPA are identical, as IPA score matching is only used to refine accuracy.

English as a different language when pronunciation is inaccurate.

As shown in Table 6, when transcripts alone are used as input to the LLM, *tiny* yields better assessment results than *large-en*. This observation can be illustrated through an analogy: using *large-en* is like speaking to a listener with excellent English comprehension – they can understand you even if your pronunciation is poor. In contrast, the *tiny* model resembles a listener with limited English ability, who can only understand clearly articulated speech. Whether a person with strong English listening comprehension (i.e., *large-en*) can understand you provides less insight into your pronunciation. In contrast, if people with weaker listening ability (i.e., *tiny*) can understand you easily, it indicates that your pronunciation is good.

Although the transcripts from *tiny* models perform better on their own, the *large-en* model is more effective within the TextPA framework. In TextPA, we incorporate the IPA and CMU sequences along with the transcript. Inaccurate pronunciation can lead to unnatural IPA and CMU sequences, offering similar insights to the transcript of *tiny* model. In addition, because the transcript serve as a baseline for comparison, excessive ASR errors introduce noise that reduces reliability. Overall, we believe that a stronger ASR model, such as *large-en*, is the better choice within the TextPA structure.

	Accuracy		Fluency	
	large-en	*tiny*	*large-en*	*tiny*
MultiPA data				
LLM: all *(gpt-4o-mini)*	**0.643**	0.569	**0.650**	0.546
LLM: transcript	0.404	**0.556**	0.432	**0.442**
Speechocean				
LLM: all *(gemini-2.0-flash)*	0.456	**0.481**	**0.557**	0.523
LLM: transcript	0.313	**0.409**	0.310	**0.431**

Table 5: Impact of ASR transcription quality.

4.5 Analysis of Basic vs. Detailed Scoring Guidelines

We investigated the impact of providing different instructions to the LLM, including basic and detailed scoring guidelines. The basic scoring guideline instructs the LLM *a scoring range (1-5), where a higher score indicates better pronunciation, with a score of 5 reflecting native-speaker proficiency.* The detailed scoring guideline, on the other hand, provides the same detailed annotation guidelines used by human annotators. The detailed guidelines define the language proficiency for each score level. For example, for MultiPA data, an accuracy score of 5 means *"Excellent: The overall pronunciation is nearly perfect with accurate articulation of all sounds,"* while a score of 4 means *"Good: Minor pronunciation errors may be present, but overall, the pronunciation is highly accurate and easily understandable"*, and so on. Results suggest that the effectiveness is dataset-dependent, possibly influenced by how the guidelines are written. However, incorporating a detailed scoring guideline has the potential to reduce performance, while also lengthening the input text prompt and increasing model operating costs.

	Accuracy		Fluency	
	Basic	Detailed	Basic	Detailed
MultiPA data				
LLM: all *(gpt-4o-mini)*	**0.643**	0.500	**0.650**	0.543
LLM: all *(gemini-2.0-flash)*	0.554	**0.596**	**0.556**	0.499
Speechocean				
LLM: all *(gpt-4o-mini)*	0.420	**0.474**	0.466	**0.544**
LLM: all *(gemini-2.0-flash)*	0.456	**0.470**	0.557	**0.561**

Table 6: Performance with basic or detailed guidelines.

5 Related Work

5.1 Speech Pronunciation Assessment

Speech pronunciation assessment models can be categorized into closed- or open-response scenarios. In closed-response settings, L2 learners read a predetermined sentence, which serves as the ground-truth transcript for the model to guide the assessment. A common approach in this scenario extracted Goodness of Pronunciation (GoP) features to train an acoustic model (Gong et al., 2022; Do et al., 2023). In addition to GoP, various other features have been explored for model training, including acoustic embeddings from self-supervised learning (SSL) models, prosodic features such as duration and energy, and transcript-based features such as word embeddings (Chao et al., 2022; Yan et al., 2025). In (Wu et al., 2025), researchers fine-tuned an LLM using audio tokens and text prompts to provide feedback on phone errors. However, the performance of models trained with ground-truth transcripts may degrade significantly when such transcripts are unavailable. On the other hand, open-response scenarios allow learners to speak freely or respond to prompts, enabling a more authentic evaluation of their pronunciation skills. Models designed for open-response tasks do not rely on ground-truth transcripts. Instead, they leverage ASR outputs or avoid ASR entirely (Lin and Wang, 2021; Kim et al., 2022; Chen et al., 2024; Liu et al., 2023b). Most prior studies rely on audio-score pair data to train acoustic models for pronunciation assessment, whereas zero-shot approaches have been largely unexplored. In (Liu et al., 2023a), researchers scored pronunciation based on the number of incorrectly recovered tokens from an SSL model. However, like other previous studies, it provided only numerical feedback instead of more interpretable or explainable assessments.

5.2 LLM for Language Learning

LLMs have had a significant impact on education, with many studies exploring how tools like ChatGPT can support language learning (Lo et al., 2024; C Meniado, 2023). These models have proven effective in helping learners identify and correct writing errors, improve the quality of their writing (Barrot, 2023), and receive automated feedback (Mizumoto and Eguchi, 2023). Few studies have focused on using LLMs to support speaking skills. (Kim and Park, 2023) used ChatGPT as a conversational partner in role-playing tasks, while (Lee et al., 2023) used it to generate topics for oral practice. A study by (Wang et al., 2023) used ChatGPT to assess how well ESL learners placed pauses in their speech. However, the potential of LLMs to support other aspects of oral language skills, such as pronunciation accuracy and fluency as in TextPA, remains under-explored.

6 Conclusion

We propose TextPA, a zero-shot pronunciation assessment method that leverages interpretable, textual representations of speech signals to assess pronunciation accuracy and fluency. These descriptions include transcripts, IPA, and CMU phoneme sequences, collectively reflecting pronunciation characteristics. Specifically, semantically unnatural transcripts may signal pronunciation issues, mismatches between canonical and recognized phoneme sequences reflect articulation errors, and inappropriate pauses embedded in CMU sequences reveal disfluencies. Experimental results demonstrate that LLMs can effectively leverage textual description of speech to assess different aspects of pronunciation. Unlike conventional models trained on audio-score pairs, TextPA operates without supervision. TextPA focuses on human-readable representations and prior knowledge of pronunciation, aiming to provide interpretable and explainable feedback that go beyond a score. We hope this work offers a new perspective on pronunciation assessment. Building on our initial exploration, future research could further develop methods to more effectively integrate TextPA with audio-trained models, combining their strengths to improve assessment accuracy and feedback quality for learners.

Limitations

While prosody is an important aspect of pronunciation, we found it difficult to effectively assess using our text-based approach. Compared to accuracy and fluency, prosodic features such as rhythm and intonation are harder to describe precisely in written form, making them less suitable for methods that rely solely on textual representations. As a result, the LLM struggled to reliably evaluate prosody without compromising assessment performance on accuracy and fluency. In addition, both the LLM and the ASR system introduce variability across runs, leading to inconsistent assessment results. In addition, budget constraints limited our ability to use the most advanced LLMs or to evaluate large ALMs across all settings. Finally, although LLM's reasoning appeared reasonable in our case study, no established metric exists to automatically verify its correctness, and exhaustive manual evaluation of every sample is beyond the scope of this study. These limitations suggest future work in prosody modeling, dataset expansion, and automatic reasoning evaluation.

Although certain words may have multiple valid pronunciations depending on the speaker's accent, our study did not consider accent variation, since the majority of the data involved attempts to mimic General American English. Consequently, a potential risk of this study is an overemphasis on a single accent. While many English learners aim to emulate native speakers, the more practical goal in everyday communication is to express one's opinions clearly and be understood. This highlights the importance of balancing pronunciation assessment systems between intelligibility and nativeness. When such systems overemphasize native-like pronunciation, which is often tied to a specific accent, they might erroneously mark understandable speech as "wrong." Failing to strike this balance can marginalize learners' linguistic identities and encourage unnecessary *accent reduction* at the expense of communicative effectiveness. In addition, an overly narrow model can reinforce the idea that only a single variety of English is valid, thereby undermining the rich diversity of global English accents.

References

Jessie S Barrot. 2023. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57:100745.

Mary E Beckman and Julia Hirschberg. 1994. The tobi annotation conventions. *Ohio State University*.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Joel C Meniado. 2023. The impact of ChatGPT on english language teaching, learning, and assessment: A rapid review of literature. *Arab World English Journals (AWEJ) Volume*, 14.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2022. 3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In *Proc. APSIPA ASC 2022*, pages 575–582. IEEE.

Yu-Wen Chen, Zhou Yu, and Julia Hirschberg. 2024. MultiPA: a multi-task speech pronunciation assessment model for open response scenarios. In *Proc. INTERSPEECH 2024*, pages 297–301.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. 2024. PAM: Prompting audio-language models for audio quality assessment. In *Proc. INTERSPEECH 2024*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Hierarchical pronunciation assessment with multi-aspect attention. In *Proc. ICASSP 2023*. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *Proc. ICASSP 2023*. IEEE.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based mmulti-aspect multi-granularity non-native english speaker ppronunciation assessment. In *Proc. ICASSP 2022*, pages 7262–7266. IEEE.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic pronunciation assessment using self-supervised speech representation learning. In *Proc. INTERSPEECH 2022*, pages 1411–1415.

Sol Kim and Seon-Ho Park. 2023. Young korean EFL learners' perception of role-playing scripts: ChatGPT vs. textbooks. *Journal of English Language and Linguistics*, 23:1136–1153.

Hyungmin Lee, Chen-Chun Hsia, Aleksandr Tsoy, Sungmin Choi, Hanchao Hou, and Shiguang Ni. 2023. VisionARy: Exploratory research on contextual language learning using AR glasses with ChatGPT. In *Proceedings of the 15th biannual conference of the Italian SIGCHI chapter*, pages 1–6.

Binghuai Lin and Liyuan Wang. 2021. Deep feature transfer learning for automatic pronunciation assessment. In *Proc. INTERSPEECH 2021*, pages 4438–4442.

Binghuai Lin and Liyuan Wang. 2022. Exploiting information from native data for non-native automatic pronunciation assessment. In *Proc. SLT 2022*, pages 708–714. IEEE.

Hongfu Liu, Mingqian Shi, and Ye Wang. 2023a. Zero-shot automatic pronunciation assessment. In *Proc. INTERSPEECH 2023*, pages 1009–1013.

Wei Liu, Kaiqi Fu, Xiaohai Tian, Shuju Shi, Wei Li, Zejun Ma, and Tan Lee. 2023b. An ASR-free fluency scoring approach with self-supervised learning. In *Proc. ICASSP 2023*. IEEE.

Chung Kwan Lo, Philip Leung Ho Yu, Simin Xu, Davy Tsz Kit Ng, and Morris Siu-yung Jong. 2024. Exploring the application of ChatGPT in ESL/EFL education and related research issues: a systematic review of empirical studies. *Smart Learning Environments*, 11(1):50.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML 2023*, pages 28492–28518.

Temple F Smith, Michael S Waterman, and 1 others. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 20234. SALMONN: Towards generic hearing abilities for large language models. *Proc. ICLR 2024*.

Ke Wang, Lei He, Kun Liu, Yan Deng, Wenning Wei, and Sheng Zhao. 2025a. Exploring the potential of large multimodal models as effective alternatives for pronunciation assessment. *arXiv preprint arXiv:2503.11229*.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, and 1 others. 2025b. Enabling auditory large language models for automatic speech quality evaluation. In *Proc. ICASSP 2025*. IEEE.

Zhiyi Wang, Shaoguang Mao, Wenshan Wu, Yan Xia, Yan Deng, and Jonathan Tien. 2023. Assessing phrase break of ESL speech with pre-trained language models and large language models. In *Proc. INTERSPEECH 2023*, pages 4194–4198.

Minglin Wu, Jing Xu, Xueyuan Chen, and Helen Meng. 2025. Integrating potential pronunciations for enhanced mispronunciation detection and diagnosis ability in llms. In *Proc. ICASSP 2025*. IEEE.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. In *Proc. INTERSPEECH 2021*, pages 2113–2117.

Bi-Cheng Yan, Yi-Cheng Wang, Jiun-Ting Li, Meng-Shin Lin, Hsin-Wei Wang, Wei-Cheng Chao, and Berlin Chen. 2025. ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization. In *Proc. ICASSP 2025*. IEEE.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. Speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proc. INTERSPEECH 2021*, pages 3710–3714.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In *Proc. ICASSP 2022*, pages 8167–8171. IEEE.

A Prompt

Figure 4 shows the TextPA prompt for LLM; ALM prompt follows a similar format, but does not include input format instructions. We observed that Gemini is more likely to return results that do not match the required format, whereas GPT tends to produce outputs that can be directly saved as JSON files. If the model fails to generate a correctly formatted output for a given test sample, we re-run it until a valid result is obtained.

B Prosody assessment

We investigate whether LLM could assess prosody from textual descriptions. We only used the MultiPA data for this part of the study, as most sentences in Speechocean are short and do not contain sufficient prosodic variation for a reliable assessment. First, we prompted the LLM to evaluate prosody in addition to accuracy and fluency. As shown in Table 7, the model performs worse in terms of prosody assessment compared to fluency and accuracy. In addition, introducing prosody as an additional assessment criterion leads to a decrease in the model's performance in both accuracy and fluency.

We explore textual descriptions of prosody using annotations from the ToBI (Tones and Break Indices) system (Beckman and Hirschberg, 1994)[7],

[7] https://github.com/monikaUPF/PyToBI

You are an expert evaluator of English pronunciation. Assess the accuracy and fluency of the given text input on a scale of 1 to 5, with higher scores indicating better performance. A score of 5 represents native-speaker-level proficiency.

Input format:
{"Transcript": "<Recognized ASR sentence>",
"Phonemes_CMU": "<Recognized CMU pronouncing phoneme sequence, with (time.s pause) indicating pauses in speech.>",
"Phonemes_IPA": "<Recognized IPA pronouncing phoneme sequence.>"}

Task: Return a dictionary with the following format:
{"Accuracy": <the assessment accuracy score>,
"Fluency": <the assessment fluency score>,
"Reasoning": <detailed reasoning for the assigned score>}

Note: Do not include any other text other than the json object.
Input:

Figure 4: LLM prompt.

	Accuracy	Fluency	Prosody
LLM: all *(gpt-4o-mini)*	0.633	0.678	-
LLM_p: all *(gpt-4o-mini)*	0.590	0.549	0.243

Table 7: LLM performance with and without prosody assessment. *LLM_p: all* is *LLM: all* with the introduction of prosody as an additional assessment criterion. Note that the transcript is generated using *turbo* version of Whisper, an optimized version of *large-v3* that provides faster transcription with minimal loss in accuracy. The results indicate that *turbo* performs comparably to *large-v3-en*. (Section 4.1)

which provides a standardized approach to annotate intonation and phrasing patterns in spoken English. ToBI includes two primary components: the break index and the tone index, both of which are crucial for understanding the prosody of speech signals. The break index ranges from 0 to 4 and is defined as follows:

- 0: Clear phonetic marks for clitic groups
- 1: Most phrase-medial word boundaries
- 2: Strong disjuncture, pause or virtual pause, no tonal marks
- 3: Intermediate intonation phrase boundary
- 4: Full intonation phrase boundary

The tone index includes the following categories:

- H: High pitch in the local pitch range
- L: Low pitch in the local pitch range
- ∗: Pitch accent, indicating that the word is stressed
- %: The end of an intonation phrase
- \- or −−: A phrase's accent

Table 9 presents a selection of examples from our attempts to assess prosody using an LLM. The experimental results indicate that the LLM is less effective in assessing prosody, and requiring it to do so leads to a decline performance in accuracy and fluency. A possible reason for this is that prosody is harder to capture accurately using textual descriptions. Since prosody is less commonly expressed in written form, the LLM has more difficulty leveraging its inherent knowledge for prosody assessment.

C Reasoning styles of GPT and Gemini

We compare the reasoning styles from GPT-4o-mini and Gemini-2.0-Flash models. We find that Gemini is more likely to cite the provided CMU and IPA phoneme sequences in its responses to support its assessment score. In contrast, GPT tends to provide more generalized observations and is more likely to suggest directions for improvement. Table 8 presents example reasoning outputs from both models given the same input.

gpt-4o-mini	The accuracy score of 4 reflects a strong understanding and correct pronunciation of most words, with only a few minor errors in phoneme realization (e.g., 'thought' was rendered correctly, but some phonemes appeared less fluid). The fluency score of 3 indicates that while the speech is coherent, there are noticeable pauses and slight hesitations that disrupt the overall flow. The transitions between phrases can be improved for a more natural delivery. Overall, the speaker demonstrates good pronunciation and comprehension but could work on smoother delivery.
gemini-2.0-flash	The accuracy is reasonable, with some mispronunciations, such as 'TH IH NG K' for 'think' instead of 'θ ɪ ŋ k', 'S OW T CH ER N AE T R AH L' for social network. However, most words are understandable. The fluency is moderate. There are several pauses which affect the natural flow of speech. There is also some word repetition (**ER R ER R**).

Table 8: LLM Reasoning Comparison.

Index	Accuracy	Fluency	Prosody	Prompt	Textual description of prosody
LLM$_A$	0.467	0.561	0.294	ToBI_sequence": "<Recognized ToBI sequence.>	"L-L% !H* L-L% L* L* H*+L L+H* L-H% L+H* L* L* L-L% L* H* L* L*+H L-H% H*+L L* L-L% H-L% L-L% L* H* H-L% L* L*+H LH- L*" **(Note: raw ToBI tone indices.)**
LLM$_B$	0.545	0.500	0.172	"Prosody_annotated_text": "<Sequence of ASR-recognized words with prosodic labels. '*' indicates a pitch accent, and '%' indicates a phrase boundary. Labels appear in parentheses after the corresponding word."	"depends (%) i mean it depends (*, %) on (*) what (*) i'm looking (*) for (*, %) if i'm (*) going to buy (*, %) like (*) a phone or (*) computer (*, %) i would definitely (*) choose big ones (*, %) because (%) the (%) quality (*) of the product (%) is more (*) reliable (*, --) for sure (*)" **(Note: Simplified ToBI tone indices, including pitch accents, phrase accents, and boundary tones, are provided along with the corresponding words in the transcript.)**
LLM$_C$	0.494	0.617	0.231	"Prosody_annotated_text": "<Sequence of ASR-recognized words with prosodic labels. '*' indicates a pitch accent, '--' indicates a phrase accent, and '%' indicates a phrase boundary. Labels appear in parentheses after the corresponding word."	"depends (%). i mean it depends (*). on (*) what (*) i'm looking, for (*). if i'm (*) going to buy (*). like (*) a phone or (*) computer. i would definitely, choose big ones (*). because (%). the (%). quality (*) of the product (%). is more (*) reliable, for sure (*)" **(Note: Simplified ToBI tone indices are used. Only the final tone index for each word is considered.)**
LLM$_D$	0.593	0.604	0.353	"Prosody_annotated_text": "<Sequence of ASR-recognized words with prosodic labels. '*' indicates a pitch accent, '--' indicates a phrase accent, and '%' indicates a phrase boundary. Labels appear in parentheses after the corresponding word."	"depends (--,%).... i mean it depends (*).... on (*) what (*) i'm looking (*).. for (*).... if i'm (*) going to buy (*).... like (*) a phone or (*) computer (*).... i would definitely (*).. choose big ones (*).... because (--,%).... the (--,%).... quality (*) of the product (--,%).... is more (*) reliable (*)... for sure (*)" **(Note: Simplified ToBI tone indices are used. Break index information is represented by the number of dots, with more dots ("....") indicating a longer break.)**
LLM$_E$	0.539	0.680	0.3043	"Transcript_prosody": "<Sequence of ASR recognized word with prosody information.>"	"depends ....i mean it depends ....on what i'm looking ..for ....if i'm going to buy ....like a phone or computer ....i would definitely ..choose big ones ....because ....the ....quality of the product ....is more reliable ...for sure"

Table 9: LLM performance in the presence of textual prosody descriptions. The Prompt column displays the additional instructions given to the LLM, beyond the standard prompt shown in Figure 4. The Textual Description of Prosody column illustrates an example input provided to the LLM.