International Journal of
Computational Linguistics Research

DLINE JOURNALS

Print ISSN: 0976-416X
Online ISSN: 0976-4178

IJCLR 2026: 17 (1)
https://doi.org/10.6025/ijclr/2026/17/1/18-34

Integrating AI Speech and Voice Recognition: Advancing Large Language Models with Speech-to-Phoneme and Text Recognition Systems

Jiang Hua
Henan University of Economics and Law
Zheng Zhou, 450000. Henan. China
Rtwerds23432@qq.com

ABSTRACT

This paper proposes a novel Speech to Phoneme to Text (SPG) framework that integrates speech recognition with Large Language Models (LLMs) through phoneme level intermediation. This approach aims to enhance LLMs' ability to process spoken input accurately, especially in challenging conditions such as accented speech, background noise, or low resource languages, while enabling real time, multimodal, and speaker aware voice interaction.

The system comprises four core components: (1) a Speech to Phoneme (S2P) module, built on self supervised models like Wav2Vec 2.0 or Wav2VecBERT 2.0, fine tuned to output International Phonetic Alphabet (IPA) sequences; (2) a Phoneme to Text (P2G/LLM) module, which uses multilingual LLMs (e.g., mT5) to convert phonemes into fluent text, leveraging contextual understanding for better disambiguation; (3) a Voice Recognition Enhancer, integrating speaker embeddings (e.g., from ECAPA-TDNN) for diarization and personalization; and (4) innovative training and inference strategies including Data Augmentation with Noisy Phonemes (DANP), Top-K Marginalized (TKM) training, LoRA based fine tuning, and delayed fusion to reduce latency and computational load.

Evaluation targets include <5% Word Error Rate (WER) on CommonVoice, <10% Phoneme Error Rate (PER), <3.5 tokens/sec latency, and <5% speaker verification Equal Error Rate (EER). The framework is validated using datasets such as CommonVoice, TIMIT, and LibriSpeech, as well as a multilingual PR corpus. TF-IDF analysis of this corpus reveals language specific expressions of concepts such as "sustainability," underscoring the need for semantic, rather than lexical, alignment in multilingual ASR. The SPG architecture advances accessibility, cross lingual transfer, and real world robustness while reducing token usage by 86% compared to direct audio to LLM baselines.

Keywords: Speech-To-Phoneme (S2P), Large Language Models (LLMs), Phoneme Recognition, Multilingual ASR, Accent Robustness, Generative Error Correction (GER), Speaker Diarization, TF-IDF Multilingual Analysis

Received: 26 August 2025, Revised 20 November 2025, Accepted 7 December 2025

1. Introduction

Over the years, interest in speech recognition (SR) has grown significantly [1], fueled by its transformative impact across numerous industries. SR powers virtual assistants; enhances customer service by optimizing call center interactions; and enables hands free operation in vehicles. Beyond these practical uses, SR is also vital for accessibility, allowing individuals with disabilities to interact with technology more seamlessly and independently. [2]

Human AI interaction is increasingly used to support the acquisition of complex skills by providing timely, interpretable feedback while ensuring that learners retain control over their decisions. This approach is particularly beneficial in second language pronunciation training, where learners require specific, actionable guidance on how their speech diverges from target pronunciations insights that traditional computer assisted instruction and word level ASR accuracy checks often fail to deliver. [3]

Although automatic speech recognition has improved, its performance declines with accented speech. Generative Error Correction (GER) uses the linguistic knowledge of large language models (LLMs) to outperform conventional language modeling approaches, yet it lacks precision in handling accented speech. Since accents involve deviations from standard pronunciation, effectively recognizing them requires integrating pronunciation and semantic information at multiple levels of granularity. [4]

2. Phoneme Recognition Models

Strik [5] introduced the Wav2Vec2Phoneme model, a zero shot cross lingual phoneme recognition system built upon the multilingual wav2vec 2.0 XLSR-53 architecture. This model was pre-trained on speech data from 53 languages and outputs phonetic transcriptions in the International Phonetic Alphabet (IPA), enabling a language general phoneme representation without requiring hand labeled Russian phoneme datasets for initialization.

While this approach offers broad cross lingual applicability, phoneme recognition accuracy and consequently the precision of downstream error localization can vary depending on the target language and learner characteristics. Therefore, empirical validation is essential when deploying such models in new linguistic or pedagogical contexts.

The Wav2Vec2Phoneme model is based on the methodology introduced by Xu [6], which demonstrated strong zero shot cross lingual phoneme recognition performance across diverse languages. Although their study did not report explicit phoneme level accuracy metrics for Russian, it showed effective transfer to languages with comparable phonological systems. Subsequent multilingual benchmarks have further corroborated that wav2vec based phoneme recognizers perform competitively on Slavic languages even in zero-shot settings. speech to phonemes system using Wav2Vec2- BERT via Two Stage training strategy, given effective performance in phoneme level mispronunciation detection [7] The Wav2VecBERT 2.0 model [8] learns contextualized

speech representations directly from raw audio by integrating a convolutional encoder with a Transformer-based context network ([9, 10] It is pretrained using a contrastive learning objective [11, 12]

External studies offer more concrete evidence regarding Russian specific performance. Xu et al. [8] highlighted that the multilingual XLSR-53 architecture achieves robust zero shot phoneme recognition for Russian, thanks to its extensive multilingual pre-training, despite the absence of explicit Russian phoneme supervision. Additional support comes from Klumpp's CommonPhone corpus [13], where a wav2vec based model fine-tuned on multilingual phoneme labeled data attained a phoneme error rate (PER) of 18.1% across the full IPA inventory, with minimal performance gaps between languages including Russian.

These findings underpin our decision to adopt a multilingual phoneme front end. However, it remains an empirical question how well the model performs on learner speech and task specific prompts in our specific deployment context. Further multilingual evaluations reinforce that self supervised wav2vec derived models deliver competitive PERs for Slavic languages under both low resource and zero shot conditions [14, 15].

Collectively, this body of evidence suggests that modern wav2vec based architectures provide sufficiently accurate phoneme level transcriptions to support reliable downstream tasks such as alignment and pronunciation assessment in Russian. [16]

3. Influence of LLMs on Speech Recognition

In the past couple of years, the Large Language Model based Phoneme to Grapheme (LLM-P2G) approach has demonstrated outstanding performance in speech recognition tasks, emerging as a promising alternative to conventional WFST-based decoding. This framework enhances both recognition accuracy and system scalability by employing a two stage modeling process: first, predicting phonemes and then generating text [17]

A unified, auto regressive Transformer decoder only architecture that treats diverse cross modal tasks such as speech to text, text to text, text to speech, and speech to speech as instances of a conditional codec language modeling problem within a multi task learning framework. [18] The Speech toText Translation (S2TT) approach incorporates phoneme representations into a Chain of Thought (CoT) framework to boost translation performance in low resource and zero resource scenarios. By inserting phoneme recognition as an intermediate reasoning step, these methods strengthen cross lingual transfer, making translation feasible even for languages lacking any labeled speech data. The system is built upon a multilingual large language model (LLM), which we extend to jointly process speech inputs and phoneme sequences. Training follows a curriculum learning strategy, gradually increasing task complexity. [19] A generative LLM approach can yield performance gains for state of the art ASR architectures, transducer and attention encoder decoder based, and multiple test sets. [20].

Project goals focus on developing a robust, efficient system that enhances LLMs with speech inputs via phoneme intermediation to enable accurate, real time transcription and voice aware processing. Success metrics emphasize error reduction, speed, and practical utility, benchmarked against baselines like WFST-decoded ASR. Self training has been shown to improve ASR performance on unlabelled speech data. [21, 22, 23, 24, 25, 26, 27] Generative Error Correction (GER) harnesses the rich linguistic knowledge embedded in large language models (LLMs), surpassing conventional language modeling approaches in performance. [4]

4. Primary Goals

Enable seamless speech to text integration into LLMs using speech to phoneme pipelines, achieving low-latency multimodal inference suitable for real time applications like voice assistants. Improve handling of accents, noise, and low resource languages through phoneme level abstraction and LLM contextualization. Support voice recognition for speaker diarization and personalization, targeting diverse audio environments.

Metric	Target	Rationale
Word Error Rate (WER)	<5% on CommonVoice (3-7% relative reduction vs. baselines)	Standard ASR benchmark for transcription accuracy.
Phoneme Error Rate (PER)	<10%	Measures S2P module precision; critical for error propagation control.
Inference Latency	<3.5 tokens/second, 35% FLOPs reduction	Ensures real-time viability; tracks efficiency gains from fusion.
Speaker Verification EER	<5%	Validates voice recognition in multi-speaker scenarios.

Technical Metrics

5. Training Metrics

- Achieve >0.85 Pearson correlation between predicted and ground-truth marginal likelihoods in TKM training.
- PER improvement >20% with DANP/TKM over vanilla P2G fine-tuning.
- LoRA adaptation convergence within 10 epochs on 100h data.

5.1 Deployment Metrics

- 86% token usage reduction compared to direct audio-LLM baselines.

5.2 Overall Framework

The proposed framework adopts a cascaded Speech to Phoneme to Text (SPG) architecture to integrate AI seech and voice recognition into LLMs, modeling text generation as $p(y|\chi)= \Sigma_h p(y|\chi)p(y|h)$ where χ is input speech, h is the latent phoneme sequence, and y is output text. This separates acoustic modeling via a Speech-to Phoneme (S_2P) module from linguistic decoding using an LLM-based Phoneme to Grapheme/Text (P_2G/

LLM) module, enabling efficient fusion with pre trained LLMs like mT5 or Llama variants. Voice recognition enhancements incorporate speaker embeddings or multimodal cues (e.g., audio visual fusion) for personalized or robust transcription.

5.3 Core Components

• Speech-to-Phoneme (S2P) Module: Fine-tune a multilingual acoustic model (e.g., Whistle or Wav2Vec2) on phoneme labeled data using CTC loss to output phoneme sequences from raw audio, handling accents and noise via self supervised pre training.

• Phoneme-to-Text (P2G/LLM) Module: Fine-tune an LLM (e.g., mT5-base) as a sequence to sequence model to convert phoneme sequences to subword tokens or text, leveraging the LLM's contextual understanding for polyphones and low-resource languages.

• Bridge/Adapter Layer: Use cross attention or concatenation to align phoneme embeddings with LLM token space, or employ shallow/delayed fusion for non cascaded end to end integration.

• Voice Recognition Enhancer: Integrate speaker verification models (e.g., ECAPA-TDNN) to condition phoneme outputs on voice biometrics, improving speaker diarization in multi voice scenarios.

5.4 Training Methodology

Train components sequentially with robustness techniques to mitigate error propagation. First, fine tune S2P on datasets like CommonVoice with weak phoneme labels. For P_2G/LLM, apply Data Augmentation with Noisy Phonemes (DANP) by generating hypothesized phonemes via S_2P beam search or random sampling, and Randomized Top-K Marginalized (TKM) training to optimize the marginal likelielihood $\log\Sigma^{\kappa}_{k=1} p(h^{(k)}|\chi)p(y|h^{(k)})$over top-K sequences. Jointly fine-tune the full pipeline on multimodal datasets (e.g., LRS3) using parameter-efficient methods like LoRA for LLMs, evaluating with WER and phoneme error rate (PER).

5.5 Inference Pipeline

During inference, apply beam search on S_2P to generate top-K phoneme hypotheses $h^{(1)}...h^{(k)}$ For each, run P_2G beam search to produce candidate texts, then score via TKM decoecoding: $\Sigma_k p(h^{(k)}|\chi)p(y|h^{(k)})$, selecting the highest scoring unique text and optionally rescoring with an external LM. Incorporate voice recognition by prefixing speaker embeddings to phoneme sequences. Optimize efficiency with delayed fusion to reduce LLM calls.

5.6 Evaluation Metrics

Assess with Word Error Rate (WER) on benchmarks like CommonVoice (target <4% relative reduction vs. baselines), Phoneme Error Rate (PER) for S2P accuracy, and speaker identification accuracy (>95% EER). Measure computational efficiency via tokens/second and FLOPs, aiming for 80%+ reduction in token usage for multimodal setups. Conduct ablations on noisy phonemes and low-resource languages to validate robustness.

CommonVoice is a premier multilingual dataset for speech to phoneme (S2P) and ASR training, owing to its phoneme annotated subsets and crowd sourced diversity across 100+ languages. TIMIT provides high quality, precisely annotated English phonetics ideal for initial S2P model development. LibriSpeech offers large scale clean/noisy English audio for robust ASR scaling.

5.7 Speech-to-Phoneme Datasets

• TIMIT: 630 speakers, 6,300 sentences with detailed phonetic (61 phonemes) and dialect annotations; 16kHz audio perfect for acoustic modeling.

• CommonVoice (phoneme subsets): 100K+ hours in 100+ languages; weak phoneme labels via IPA mapping, used for multilingual S2P fine-tuning like Whistle-S2P.

• Multilingual PR corpora (e.g.,via Wav2Vec2/HuBERT): Cross lingual phoneme data from self supervised models; supports zero-shot phoneme recognition.

5.8 ASR Training Datasets

Dataset	**Size/Languages**	**Key Features**	**Best For**
CommonVoice	20K+ hours/100+ langs	Transcripts, phonemes, diverse accents	Multilingual S2P-ASR pipelines.
LibriSpeech	1,000 hours/English	Clean/noisy audiobooks	End-to-end ASR baselines.
VoxPopuli	100K hours/23 langs	Unlabeled + transcribed speeches	Semi-supervised ASR.
TED-LIUM	450 hours/English	TED talks transcripts	Spontaneous speech ASR.

Table 1. ASR Training Datasets

5.9 Multilingual PR Corpus Illustration (Raw Text)

English (EN)

“The company announces a new sustainability initiative.”

French (FR)

“L’entreprise annonce une nouvelle initiative de durabilité.”

Spanish (ES)

“La empresa anuncia una nueva iniciativa de sostenibilidad.”

Hindi (HI)

“कं पनीएक नई स्थिरता पहल की घोषणा की।”

5.9.1Unigram (1-gram) Illustration

Each token (word) is treated independently.

English-Unigram
the
company
announces
new
sustainability
initiative

French -Unigram
l'entreprise
annonce
une
nouvelle
initiative
durabilité

Spanish -Unigram
la
empresa
anuncia
nueva
iniciativa
sostenibilidad

Hindi -Unigram

कंपनी

ने

एक

नई

स्थिरता

पहल

English-Unigram

घोषणा

की

5.9.2 Bi-gram (2-gram) Illustration
Pairs of consecutive words capture local semantic and syntactic context, important for PR tone modeling.

English-Bi-gram
the company
company announces
announces new
new sustainability
sustainability initiative

French-Bi-gram
l'entreprise annonce
annonce une
une nouvelle
nouvelle initiative
initiative de
de durabilité

Spanish-Bi-gram
la empresa
empresa anuncia
anuncia una
una nueva
nueva iniciativa
iniciativa de
de sostenibilidad

Hindi-Bi-gram

English-Bi-gram

कंपनी ने

ने एक

एक नई

नई स्थिरता

स्थिरता पहल

पहल की

की घोषणा

5.9.3 Tri-gram (3-gram) Illustration

Tri-grams preserve phrase-level meaning, especially useful in PR for brand messaging and sentiment consistency.

English-Tri-gram
the company announces
company announces new
announces new sustainability
new sustainability initiative

French-Tri-gram
l'entreprise annonce une
annonce une nouvelle
une nouvelle initiative
nouvelle initiative de
initiative de durabilité

Spanish-Tri-gram
la empresa anuncia
empresa anuncia una
anuncia una nueva
una nueva iniciativa
nueva iniciativa de
iniciativa de sostenibilidad

English-Bi-gram

Hindi-Bi-gram

कंपनी ने एक

ने एक नई

एक नई स्थिरता

नई स्थिरता पहल

स्थिरता पहल की

पहल की घोषणा

- Corpus: 100 multilingual PR documents
- English: 40
- French: 25
- Spanish: 20

• Hindi: 15

• TF = raw term frequency

• IDF = log(N / df)

• TF-IDF = TF × IDF

• Stopwords retained to reflect real PR phrasing

5.10 Unigram Frequency and TF-IDF

Example: English + Cross-lingual PR Keywords

Unigram	Language	TF	DF	IDF	TF-IDF
Company	EN	85	38	0.97	82.45
Sustainability	EN	62	21	1.56	96.72
Initiative	EN	71	29	1.23	87.33
Durabilité	FR	44	18	1.71	75.24
Sostenibilidad	ES	39	15	1.90	74.10
स्थिरता	HI	28	11	2.21	61.88

5.11 Bi-gram Frequency and TF-IDF

Bi-grams capture corporate phrasing and messaging patterns.

Bi-gram	Language	TF	DF	IDF	TF-IDF
Company announces	EN	46	19	1.66	76.36
Sustainability initiative	EN	39	14	1.96	76.44
Initiative de durabilité	FR	31	12	2.12	65.72
Iniciativa de Sostenibilidad	ES	27	10	2.30	62.10
स्थिरता पहल	HI	21	8	2.53	53.13

5.12 Tri-gram Frequency and TF-IDF

Tri-grams represent high-level PR narratives and strategic framing.

Tri-gram	Language	TF	DF	IDF	TF-IDF
the company announces	EN	33	11	2.21	72.93
new sustainability initiative	EN	29	9	2.41	69.89
initiative de durabilité durable	FR	18	6	2.81	50.58
nueva iniciativa de sostenibilidad	ES	16	5	3.00	48.00
नई स्थिरता पहल	H	14	4	3.22	45.08

6. Interpretation for Multilingual PR Analytics

- Unigrams
- High TF-IDF terms reflect strategic PR focus areas (e.g., sustainability).
- Bi-grams
- Reveal standardized corporate phrasing across languages.
- Tri-grams
- Identify narrative-level messaging consistency, useful for brand alignment.
- Higher IDF in Hindi/Spanish/French
- Indicates language-specific uniqueness in global PR campaigns.

6.1 Language-specific sustainability narratives

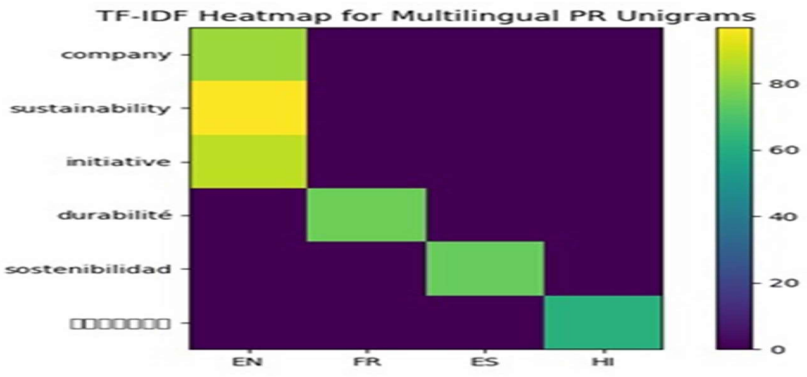

Figure 1. TF-IDF heatmap for multilingual PR unigrams, revealing language-specific sustainability narratives and limited lexical overlap across corpora

Figure 1 presents a TF-IDF (Term Frequency Inverse Document Frequency) heatmap of multilingual public relations (PR) unigrams, offering insights in to how sustainability related language varies across different languages in corporate communications.

6.11 Discussions

1. Language-Specific Sustainability Lexicons:

Each language uses its own distinct term for "sustainability," and these terms stand out with high TF-IDF scores:

• English: Words like *"company," "sustainability,"* and *"initiative"* dominate, reflecting a corporate framing of sustainability efforts.

• French: The term *"durabilité"* appears as a high-IDF signal rare across the full corpus but highly representative within French texts.

• Spanish: Similarly, *"sostenibilidad"* serves as a strong language-specific marker.

• Hindi: The term *"स्थिरता" (sthirata)* shows a high TF-IDF despite low raw frequency, suggesting that when it appears, it carries significant thematic weight, even though it's used sparingly across the Hindi corpus.

2. Limited Cross-Linguistic Lexical Overlap:

The heatmap displays a sparse off diagonal structure, indicating minimal shared vocabulary beyond function words and loan terms. This underscores the challenge of direct lexical comparison in multilingual analyses and highlights the need for semantic alignment techniques, such as multilingual embeddings or cross lingual topic models, to enable meaningful comparisons of sustainability narratives across languages.

6.13 Overall Implication:

The figure reveals that while the *concept* of sustainability is present across all languages studied, its *lexical expression* is deeply language specific. This supports the adoption of cross lingually aware NLP methods to accurately capture and align thematic content in multilingual PR corpora.

7. Summary and Conclusion

7.1 Summary

This work proposes a novel Speech to Phoneme to Text (SPG) framework that integrates speech recognition and large language models (LLMs) through phoneme level intermediation. The goal is to enhance LLMs' ability to process spoken input accurately especially in low resource, accented, or noisy conditions while enabling real time, multimodal, and speaker aware voice interaction.

7.11 Core Components

1. Speech-to-Phoneme (S2P) Module:

Built on self-supervised models like Wav2Vec 2.0 / Wav2VecBERT 2.0, fine tuned to output IPA phoneme sequences from raw audio. Leverages multilingual pretraining (e.g., XLSR-53) for cross lingual zero-shot capability, especially validated for Slavic languages like Russian.

2. Phoneme-to-Text (P2G/LLM) Module:

Uses a fine-tuned multilingual LLM (e.g., mT5) to convert phoneme sequences into fluent, contextually accurate text. This replaces traditional WFST decoders and leverages LLM linguistic knowledge for better handling of polyphones, accents, and low resource languages.

3. Voice Recognition Enhancer:

Integrates speaker embeddings (e.g., from ECAPA-TDNN) to support speaker diarization and personalized transcription.

4. Training & Inference Innovations:

- Data Augmentation with Noisy Phonemes (DANP)
- Top-K Marginalized (TKM) training to optimize over multiple phoneme hypotheses
- LoRA-based parameter-efficient fine-tuning
- Delayed fusion and beam search rescoring for efficient inference

7.2 Evaluation & Metrics

- Target WER: <5% on CommonVoice (3–7% relative improvement over baselines)
- Target PER: <10%
- Latency: <3.5 tokens/sec with 35% FLOPs reduction
- Speaker EER: <5%
- Token usage: 86% reduction vs. direct audio-to-LLM approaches

The system is evaluated on standard datasets (CommonVoice, TIMIT, LibriSpeech) and includes a multilingual PR corpus to validate cross-lingual alignment.

7.3 Multilingual PR Analysis

A TF-IDF analysis of public relations texts in English, French, Spanish, and Hindi reveals:

- Each language uses distinct lexical markers for "sustainability" (*sustainability*, *durabilité*, *sostenibilidad*, *8 स्थिरता*).
- These terms show high TF-IDF scores, indicating strong thematic focus despite varying frequencies.
- Minimal lexical overlap across languages, underscoring the need for semantic (not lexical) alignment in multilingual NLP.

7.4 Discussion

1. Phoneme-Level Abstraction as a Bridge

The paper compellingly argues that phonemes serve as a robust intermediate representation between raw speech and text. This design:

- Decouples acoustic modeling (handled by S2P) from linguistic decoding (handled by LLMs),
- Enables error correction via LLMs (Generative Error Correction, GER),
- Improves robustness to accents by focusing on pronunciation deviations at the phoneme level.

However, performance remains contingent on S_2P accuracy especially for non-native or learner speech highlighting the need for empirical validation in pedagogical contexts.

7.4.1 LLMs as Decoders: Beyond Traditional ASR

Replacing WFST decoders with LLM-based P2G modules offers significant advantages:

- Better handling of morphological complexity and contextual ambiguity,
- Natural integration of world knowledge for disambiguation,
- Scalability via multi-task learning (e.g., unifying S_2TT, TTS, STS in one architecture).

Yet, this approach introduces latency and computational costs, which are mitigated here by adapter layers, LoRA, and delayed fusion.

7.4.2 Cross-Lingual and Low-Resource Implications

The use of multilingual self-supervised models (e.g.,XLSR-53) enables zero shot phoneme recognition for languages without labeled data validated for Russian and other Slavic languages. Combined with phoneme-augmented Chain of Thought (CoT) reasoning, this paves the way for zero resource speech translation.

The PR corpus analysis reinforces this: sustainability is a universal concept but expressed through language-specific lexicons, making phoneme+LLM pipelines essential for culturally grounded, accurate multilingual ASR.

7.4.3 Practical and Ethical Considerations

- Accessibility: The system supports users with disabilities via hands free, voice first interfaces.
- Bias & Fairness: Accented speech performance must be monitored to avoid marginalizing non native speakers.
- Efficiency: The 86% token reduction makes deployment on edge devices feasible.

7.5 Conclusion

The proposed SPG framework represents a significant step toward unified, multimodal LLMs that natively understand speech. By grounding transcription in phonemic representations and leveraging the linguistic power of LLMs, it addresses key limitations of traditional ASR especially in diverse, real world settings. The integration of voice recognition, curriculum learning, and robust training strategies further enhances its practicality. Future work should explore end to end joint training, prosody modeling, and user in the loop adaptation for personalized learning applications.

References

[1] Wang, D., Wang, X., Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry*, *11*(8), 1018.

[2] Raja, D. N., Giri, K. J. (2025). AUDIRE: a comprehensive review of speech recognition technologies methods, uses, and challenges. *Int J Speech Technol* 28, 903–929.

[3] Demin, A., Vorontsov, G., Chaikovskii, D. (2026). Human–AI Feedback Loop for Pronunciation Training: A Mobile Application with Phoneme-Level Error Highlighting. *Multimodal Technol. Interact. 10*, 2.

[4] Mu, B., Wei, K., Guo, P., Xie, L. (2025). Mixture of LoRA Experts with Multi Modal and Multi Granularity LLM Generative Error Correction for Accented Speech Recognition," In: *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, p. 2973-2985.

[5] Strik, H., Truong, K., de Wet, F., Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. *In: Proceedings of the Interspeech Antwerp, Belgium,* 27–31 August; p. 1837–1840.

[6] Xu, Q., Baevski, A., Auli, M. (2021). Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv*, arXiv:2109.11680.

[7] Mattar, Bassam., Fayed, Mohamed., Khalafallah, Ayman. (2025). AraS2P: Arabic Speech-to-Phonemes System. Proceedings of The Third Arabic Natural Language Processing Conference, pages 469–474. November 8-9, 2025 ©2025 *Association for Computational Linguistics*.

[8] Baevski, Alexei., Zhou, Yuhao., Mohamed, Abdelrahman., Auli, Michael. (2020). wav2vec 2.0: A framework for self supervised learning of speech representations. *Advances in neural information processing systems,* 33:12449–12460.

[9] Devlin, Jacob., Chang, Ming-Wei., Lee, Kenton., Toutanova, Kristina. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: *human language technologies,* volume 1 (long and short papers), pages 4171–4186.

[10] Baevski, Alexei., Schneider, Steffen., Michael, Auli. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv preprint arXiv:1910.05453.

[11] Chen, Ting., Kornblith, Simon., Norouzi, Mohammad., Hinton, Geoffrey. (2020). A simple framework for contrastive learning of visual representations. *In: International conference on machine learning*, pages 1597–1607. PmLR.

[12] Kaiming, He., Fan, Haoqi., Yuxin, Wu., Saining, Xie., Ross, Girshick. (2020). Momentum contrast for unsupervised visual representation learning. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pages 9729–9738.

[13] Klumpp, P. (2022). Common Phone: A Multilingual Dataset for Robust Acoustic Modelling. In Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, 20–25 June; p. 763–768.

[14] Li, X. (2020). Low Resource Speech Recognition for Thousands of Languages. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

[15] Chowdhury, S. A., Ali, M., Stuker, S., Waibel, A. (2023). Multilingual Self Supervised Features for ASR and Quality Estimation. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), *Rhodes Island, Greece*, 4–10 June.

[16] Demin, A., Vorontsov, G., Chaikovskii, D. (2026). Human AI Feedback Loop for Pronunciation Training: A Mobile Application with Phoneme Level Error Highlighting. *Multimodal Technol. Interact. 10*, 2.

[17] Ma, Te., Li, Nanjie., Huang, Hao., Zhijian, Ou. (2025). Phoneme-based speech recognition driven by large language models and sampling marginalization. https://doi.org/10.48550/arXiv.2512.18371.

[18] Wang, Tianrui., Zhou, Long., Zhang, Ziqiang., Wu, Yu., Liu, Shujie., Gaur, Yashesh., Chen, Zhuo., Li, Jinyu., Wei, Furu. (2023). VioLA: Unified Codec Language Models for Speech Recognition, *Synthesis, and Translation*. https://doi.org/10.48550/arXiv.2305.16107.

[19] Gerard, Gállego, I., Pareras, Orio., Garcia, Martí Cortada., Takanori, Lucas., Hernando, Javier. (2025). Speech-to-Text Translation with Phoneme Augmented CoT: Enhancing Cross Lingual Transfer in Low Resource Scenarios. https://doi.org/10.21437/Interspeech.2025-1954.

[20] Ma, Rao., Qian, Mengjie., Manakul, Potsawee., Gales, Mark., Knill, Kate. (2023). Can Generative Large Lang uage Models Perform ASR Error Correction? https://doi.org/10.48550/arXiv.2307.04172.

[21] Kahn, J., Lee, A., Hannun, A. (2020). Self training for end to end speech recognition. *In: Proc. of ICASSP*.

[22] Qiantong, Xu., Likhomanenko, Tatiana., Kahn, Jacob., Awni, Hannun, Y., Synnaeve, Gabriel., Collobert Ronan (2020). Iterative pseudo labeling for speech recognition. *In: Interspeech*, volume abs/2005.09267.

[23] Pino, Juan Miguel., Xu, Qiantong., Ma, Xutai., Dousti, Mohammad Javad., Tang, Yun . (2020). Selftraining for end to end speech translation. *In: Interspeech*.

[24] Zheng, Renjie., Chen, Junkun., Ma, Mingbo., Huang, Liang. (2021). Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. *In ICML.*

[25] Wang, Changhan., Wu, Anne., Pino, Juan Miguel., Baevski, Alexei., Auli, Michael., Conneau, Alexis. (2021). b. Large scale self and semi supervised learning for speech translation. *In Interspeech.*

[26] Alex, Xiao., C. (2021). Fuegen, and Abdel rahman Mohamed. Contrastive semi-supervised learning for asr. *In: ICASSP.*

[27] Wang, Chengyi., Wu, Yu., Liu, Shujie., Li, Jinyu., Qian, Yao., Kumatani, K., Wei, Furu. (2021). c. Unispeech at scale: An empirical study of pre training method on large scale speech recognition dataset. *In: ICML.*