Leveraging Intelligent Speech Training to Elevate Phonetic Accuracy and Prosodic Fluency in English Learners

Dr. Amit Khapekar[1], Dr. Nidhi Mishra[2], Vijaya Lakshmi Mandava[3], Dr. T K Rama Krishna Rao[4],
Dr. Bhuvaneswari Pagidipati[5], Dr. Prasad Devarasetty[6], Elangovan Muniyandy[7]
Assistant Professor, Department: Applied Mathematics and Humanities,
Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India[1]
Associate Professor, Department of Humanities and Sciences,
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India[2]
Associate Professor of English, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India[3]
Professor, Department of Computer Science and Engineering,
Koneru Lakshmaih Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[4]
Associate Professor of English, Dept. of English and Foreign Languages,
Sagi Rama Krishnam Raju Engineering College (A), Bhimavaram – 534204, West Godavari Dt, Andhra Pradesh, India[5]
Department of Computer Science and Engineering, DVR & Dr HS MIC College of Technology,
Kanchikacherla, Andhra Pradesh, India[6]
Department of Biosciences, Saveetha School of Engineering-Saveetha Institute of Medical and Technical Sciences,
Chennai - 602 105, India[7]

Abstract*—The successful teaching of pronunciation, as well as prosody, is the significant challenge that still remains to the English as Foreign Learning (EFL) students. Traditional pedagogical theories tend to focus on segmental phoneme accuracy but ignore suprasegmental components (stress or rhythm and intonation) which are natural and intelligible speech components. The currently available systems of computer-assisted pronunciation training (CAPT) are useful, but limited by the fact that they are based on limited acoustic models and incomplete coverage of prosodic characteristics, leading to less than optimal accuracy and limited pedagogical suitability. To overcome these shortcomings, the current paper proposes Attention-Guided Cross-Lingual Self-Supervised Learning (AG-CLSSL), a new model that is both able to combine phoneme-level representations of XLS-R (wav2vec2-large-xlsr-53) and prosodic representations of the pitch, energy, and duration through a Phoneme-Prosody Cross-Attention Fusion (PP-CAF) process. This conglomeration allows the joint and context specific representation of the speech that is further refined by the multi-task Transformer-based scoring model to jointly assess the accuracy of pronunciation, the consistency of the prosody and the general intelligibility. The framework is implemented in Python, with support of PyTorch and Hugging Face Transformers and is trained on an evaluated corpus of EFL learner speech (n=100) with a variety of L1 backgrounds, including Mandarin, Hindi, and Spanish. Experimental assessments indicate significant performance improvement with 55.4% decrease in Phoneme Error rate, 52.0 percent decrease in Word Error rate, 43.3 percent increase in Stress Placement Accuracy and 34.9 percent increase in Pitch Alignment Score. The total acoustic similarity to native speech went up by 36.1, which demonstrates the ability of AG-CLSSL to progress articulatory accuracy as well as the naturalness of prosody and provide interpretable and attention-directed information on scalable AI-based pronunciation and prosody training.

Keywords*—Automatic speech recognition; pronunciation and prosody; transformer-based phoneme identification; prosody assessment; adaptive learning algorithm

I. Introduction

For EFL learners, achieving precise pronunciation and fluid prosody is an ongoing difficult task. Where explicit vocabulary and grammar instruction tends to dominate the language learning process, pronunciation misunderstandings--through mispronunciation or inappropriate stress or intonation patterns can drastically lower intelligibility and communicative competence [1]. While phonetic instruction can take place in the classroom context, feedback is limited by the instructor's inability to provide feedback as quickly as their students will mispronounce or improperly produce sounds and suprasegmental features [2]. Furthermore, many of the class-based phonetic instructional sounds, like feedback, is put on various types of manual corrections that depend on the instructor's time, subjectivity, and in classroom situations, lack of time [3]. It is this gap in the acquisition and instruction of pronunciation and prosody that has inspired many language educators to consider technology-driven solutions for CAPT. Emerging technologies utilizing Automatic Speech Recognition (ASR) as the core functionality are now readily available to EFL learners and teachers to improve the assessment of learners' pronunciation production [4]. Pre-trained self-supervised, deep learning models like Wav2Vec2.0 and HuBERT have demonstrated excellent performance in tasks that recognize and classify phonemes, and Whisper has also further improved recognition by gaining robustness to noisy and accented speech [5]. At the same time, systems with a prosody0driven focus have integrated methodological frameworks that evaluate features of pitch and rhythm to assess learners' suprasegmental features of

speech. Even more recent iterations have specifically included gamification and reinforcement learning modalities that increase learner engagement and motivation [6], [7]. Nevertheless, there are limitations to these approaches.

Most of all existing frameworks do not take into account the typical separation between pronunciation (segmental) and prosody (suprasegmental), opting for simple concatenation of acoustic and prosodic features without modeling their interrelation [8]. This approach further neglects a critical point about intelligibility: intelligibility is determined by how segmental and suprasegmental features co-occur, making their joint modeling necessary and the previous systems often rely on multiple pre-trained backbones (e.g., Whisper, Wav2Vec2.0, MFCC pipelines), which complicate design with marginal benefits to the efficiency and interpretability [9]. Thirdly, feedback is commonly based on rule-based or reinforcement based paradigms, which are less interpretable, and groups do not necessarily develop congruent progression as individuals. [10]. In order to eliminate the drawbacks of the current models, the paper presents Attention-Guided Cross-Lingual Self-Supervised Learning (AG-CLSSL), a new model based on the usage of XLS-R embeddings combined with the use of prosodic cues to form the Phoneme2Prosody Cross-Attention Fusion (PP-CAF) layer. The architecture is able to score pronunciation and prosody context-sensitively and in multi dimensions through synergistic alignment of the phonemic representations with the aspects of pitch, duration and energy by the guidance of attention-based fusion. Multi task Transformer is used to assess articulation, prosodic fidelity, and intelligibility and adaptive curriculum learning provides personalized feedback to EFL learners. This integrative design guarantees both technical strength and pedagogical faithfulness and provides a transformative avenue in the improvement of communicative competence and integrates the current CAPT approaches. The fundamental innovation is the collective phoneme-prosody interaction modeling, using the PP-CAF layer, a single and cross-lingual fusion mechanism that cannot be reached by previous CAPT systems using the mere concatenation of features.

A. Problem Statement

Conventional teaching to EFL learners often does not provide the individual and immediate feedback of the appropriate stress patterns and intimacy, which often results in not so natural and reasonable speech. [11]. In addition, the existing ASR devices are mainly concerned with phone-level accuracy and reject the super segmental properties that consider rhythm, stress, and procession, which are central to smooth communicative interaction. Most importantly, current systems inappropriately consider different learner profiles and do not consider the first language (L1) effects, resulting in limited utility in a multilingual context [12]. With the existing unsatisfactory performance in these areas it is urgent that an intelligent, adaptive system be developed with both full use of advanced ASR and NLP techniques and with rich and personalized practice of both pronunciation and prosody. To fill this gap, the current paper suggests a new framework Attention-Guided Cross-Lingual Self-Supervised Learning (AG-CLSSL), a cross-lingual framework that combines attention components and cross-lingual self-supervised learning. The Phoneme-Prosody Cross-Attention Fusion (PP-CAF) layer is the central part of this architecture, which simultaneously learns segmental and suprasegmental speech characteristics, allowing the context-responsive, multi-dimensional evaluation and providing adaptive and learner-focused feedback that is essential to the improvement of EFL communicative competence.

B. Research Motivation

EFL learners often have a problem with a speech that is intonable because of incorrect articulation of the phonemes and loss of proper prosody. Conventional instruction and a large number of CAPT systems provide limited, delayed or non-adaptive feedback, which limits development of the learner. Regardless of the fact that self-administered models like the XLS-R have developed the assessment of pronunciation, majority of models analyze phonemic and prosodic aspects independently without considering their interrelationship. Such restrictions highlight the importance of integrative framework that concurrently models phoneme and suprasegmental cues and provides adaptive and personalized feedback hence facilitating more efficient, context-sensitive and holistic spoken language acquisition.

C. Research Significance

The study involves an attention-directed cross-lingual self-supervised model, which learns the embedding of the phoneme and the prosodic features using a novel attention cross-fusion (PP-CAF) layer. The model offered is more accurate in pronunciation and prosody evaluation besides increasing flexibility of the learner through the provision of real time feedback on the curriculum. It is important as it helps bridge the gap between technical innovation and pedagogical need as it provides a solid device in the further development of intelligent language learning. Allowing researchers and educators to promote the growth of EFL learners in the spoken language in a systematic manner, this framework facilitates more efficient and comprehensive advancement of such competence in them.

D. Key Contribution

- Introduces an integrated framework that jointly models phoneme accuracy and prosodic features, addressing a persistent gap where existing systems treat these dimensions in isolation.
- Proposes the PP-CAF layer, enabling dynamic alignment between segmental and suprasegmental features for more natural speech representation.
- Utilizes large-scale pre-trained speech embeddings to capture multilingual phonetic variations, ensuring robustness across diverse learner backgrounds.
- Employs a multi-task scoring model that simultaneously assesses phoneme error reduction, prosodic alignment, rhythm consistency, stress accuracy, and speech naturalness.
- Validated the system through Python-based implementation using Whisper and Wav2Vec 2.0, with expert evaluations confirming pedagogical reliability and real-world scalability for EFL learners.

- It offers a single CAPT system that has adaptive feedback and empirical validated gains on methodological, pedagogical, and empirical levels.

The remaining sections are organized as follows, Section II presents the literature review, Section III outlines the problem statement, Section IV discusses the results and Section V provides the conclusion and future directions.

II. Literature Review

The field of language learning has been interested in ASR as it is able to provide real time feedback regarding pronunciation and fluency. Typically, early ASR applications were developed around the speech-to-text transcription, but the recent developments enable their integration into CALL systems [13]. It is well established through studies that with ASR based pronunciation training, the learner gets objective immediate feedback that doesn't need constant human supervision; hence the learner gets to further exercise their autonomy. ASR can be applied for commercial applications such as Duolingo, Speech Ace, and Google's ELSA Speak to assess phoneme production but their feedback mechanisms are typically simple and even binary. Previous research indicates that ASR-based training is superior when learners are taught of segmental and suprasegmentally analysis [14]. Yet error detection in non-native speech remains a problem, and it requires sophisticated machine learning models that have been trained on a variety of accents and speech patterns. Phoneme recognition and error analysis have greatly been improved by incorporating NLP and deep learning. Transformer based models such as wav2vec 2.0 and Whisper use self-supervised learning to detect deviation from pronunciation without using large labelled dataset [15].

Using NLP techniques such as phoneme embeddings and forced alignment have more granular and helpful automated feedback at a phoneme boundary level. The resulting encouraging accuracy is mainly due to the contrastive learning frameworks designed to distinguish between native and non-native phoneme production. Nevertheless, one issue: Most of ASR systems have been optimized for the native speakers, which will not work very well for the L1 learners with their patterns of substituted or deleted phoneme. It is found that adaptive ASR models trained on various non-native speech corpora are needed to better phoneme-level analysis for EFL learners [16]. However, prosody learning and feedback mechanism are also important in acquiring second language. Speech naturalness and intelligibility are contributed to by stress, rhythm and intonation, and these are often overlooked when teaching pronunciation. The most common prosody errors affecting listener comprehension of a talking head or speaker confidence include incorrect stress placement and unnatural pitch contours. Segmental features serve as a focus of traditional phonetic training, with the aspects of suprasegmentals receiving less attention. It has also been shown in recent studies, that deep learning-based prosody analysis using LSTM networks is used to assess intonation patterns, speech duration and stress placement [17]. When learners receive visual and auditory cues, prosody feedback improves communicative effectiveness. Yet the work in the area of prosody evaluation based on ASR is still in low accuracy in terms of tone and stress detection, but deep learning architectures need to be stronger in comparing with native speech models [18]. The current speech learning technologies, such as rule based speech analyzers and statistical ASR models, have not been very effective as they depend on the predefined phonetic rules. HMMs and GMMs remain dominant for pronunciation scoring, while such errors cannot be captured with efficiency by these classical pattern formers [19]. When compared with deep learning-based ASR models especially, end to end neural networks, have shown to have higher accuracies for detecting phoneme level and prosodic deviation. However, in the existing ASR-driven learning tools, it offers only generic pronunciation score [20] and does not target at individualized learner challenges.

Besides, current commercial ASR applications based on fixed threshold error classification, which are the prevailing methods in commercial applications, still rely on false positives and the absence of any correction strategies [21]. A major limitation of existing pronunciation training systems is the lack of context aware feedback mechanisms and adaptive learning pathways. L1 interference plays a rather crucial role in EFL pronunciation training due to its negative interference of the learner's native language to English articulation. Using generalized pronunciation training is ineffective as the phoneme substitution, deletion and insertion patterns vary based on the learner's linguistic background [22]. Let's say, Mandarin speakers get stuck on English consonant clusters and Japanese speakers usually insert some vowel sounds in order to break them. Phoneme transfer errors in these cases are too difficult to learn with generic pronunciation models, which fail to account for L1 specific weaknesses. Furthermore, motivation is an important factor to take into consideration in order to pronounce, as most of the EFL learners are nervous and afraid when they are in a situation where they are supposed to speak. Reinforcement learning driven difficulty adjustments help sustaining motivation, and gamification enhances learner engagement through learning with ASR in studies [23].

While there has been longitudinal research examining the effect of ASR based pronunciation feedback on long term fluency development, there has been no longitudinal research looking at the long-term impact of ASR based pronunciation feedback on fluency development and retention. Due to these gaps in the current available pronunciation training methodologies, this study investigates NLP driven ASR frameworks allowing for fine-tuned speech models, phoneme error clustering and prosody feedback mechanisms to provide tailored pronunciation correction for EFL learners. By integrating transformer-based speech models, LSTM based prosody evaluation, and adaptive reinforcement learning, this research fills the gap between automated pronunciation training and human like corrective feedback.

The critical points in sustaining interest amongst learners during the context of ASR-based training of pronunciation, dropout may be defined as the case where the learner will discontinue training before a training proficiency level will be attained, whilst fatigue detecting may be considered as the case whereby the learner will identify signs of cognitive overload, or, the event of lack of interest [24]. Such issues can significantly harm the learning outcomes in the scenario of practicing pronouncing lessons repetitively and over a long period of time. The reinforcement learning (RL) models monitor the number of

sessions one has attended, on average, the length of each session, the number of errors and the number of retries to detect the early signs of fatigue or the absence of interest. The learners who demonstrate brief sessions, high error rates, and fewer retries can be marked as fatigued, and adjustments would consider shortening exercises or some form of gamification or motivation. In the same way, the predictive analysis of dropout can trigger proactive actions, which may be push notifications, personalized feedback, or encouraging feedback, to maintain the learners prior to the disengagement. Using various methods to dynamically increase or decrease task difficulty and reward behavior, the system facilitates individual feedback depending on the amount of attention given in real-time. The AI-based technology facilitates motivation on a long-term level and encourages learners to maintain their progress until mastering the pronunciation process.

Existing ASR-based CAPT systems are at the forefront of developing recognition of phonemes, analysis of prosody, and adaptive feedback, but they do not combine segmental and suprasegmental properties and use generic scoring and not focusing on errors peculiar to L1. This research proposes a unified framework that combines the relationships between phoneme and prosody, and it adapts itself based on the tendencies of multilingual learners and presents a completely different, context-driven method to assessing pronunciation and prosody.

III. Proposed Attention-Guided Cross Lingual Self Supervised Learning Framework for Pronunciation and Prosody Enhancement

The Attention-Guided Transformer Cross-Lingual Self-Supervised Learning (AG-CLSSL) framework proposed in this study embodies a unified approach to the concurrent modeling of pronunciation and prosody within a single architecture. The methodology commences with rigorous data collection and preprocessing using the Speech Accent Archive (SAA), which provides a standardized corpus of audio recordings. All samples are meticulously resampled, normalized, and phoneme-aligned, ensuring homogeneity across speakers and establishing a consistent basis for downstream modeling. Feature extraction is performed via XLS-R, a self-supervised speech representation model that generates contextualized phoneme embeddings grounded in both acoustic and linguistic information. These embeddings are temporally synchronized at phoneme boundaries, producing highly representative phoneme-level vectors. Simultaneously, prosodic features—including pitch, energy, and duration—are extracted and projected into a shared latent space, enabling the capture of suprasegmental dynamics. Central to the framework is the novel Phoneme–Prosody Cross-Attention Fusion (PP-CAF) layer, which facilitates bi-directional interaction between segmental embeddings and prosodic cues, allowing the model to integrate suprasegmental information into phonemic representations and thereby generate a holistic speech representation. These fused features are subsequently processed by a multi-task Transformer scoring model, which concurrently evaluates pronunciation accuracy, prosodic quality, and overall intelligibility. However, AG-CLSSL provides a different approach to combining segmental and prosodic features by explicitly learning the interaction between them in the PP-CAF layer, unlike current CAPT models where they are combined together. This is not possible with small modifications on the previous architectures since they are not aligned bi-directionally. The fact that cross-lingual self-supervised embeddings are used also makes L1 deviations useful cues, which introduces a new paradigm of adaptive multilingual assessment. The system is trained with a weighted multi-task loss function, balancing regression and classification objectives, and produces adaptive, learner-specific feedback, enabling personalized, performance-driven guidance for EFL learners. The workflow of the proposed method is illustrated in Fig. 1.

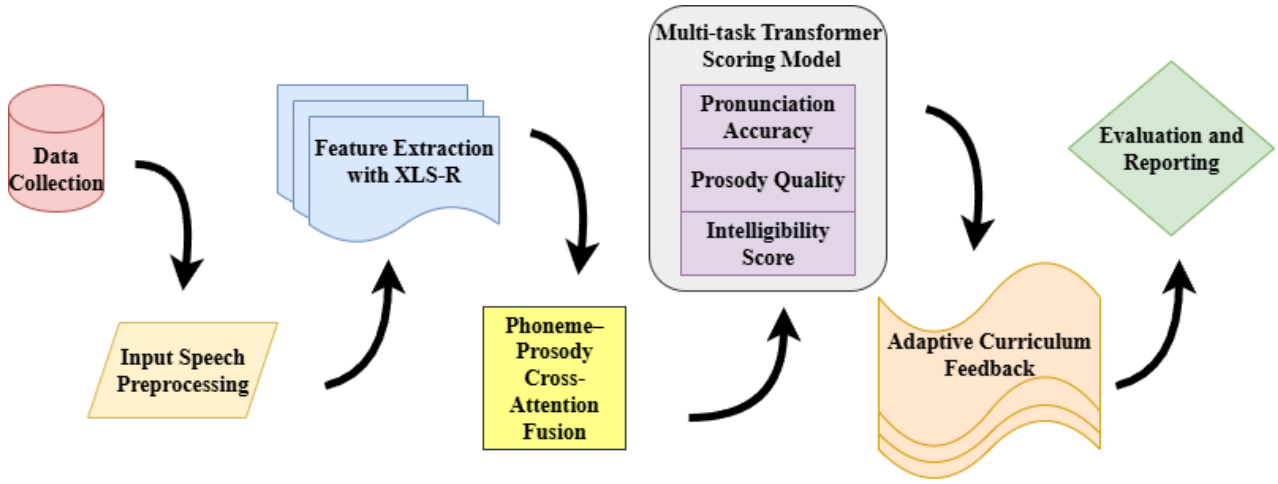

Fig. 1. Workflow of the proposed system.

Fig. 1 shows the process of the Attention-Guided Cross-Lingual Self-Supervised Learning (AG-CLSSL) model. The data in speech is collected and then preprocessed to give normal input. XLS-R is used to extract features, and results in contextual phoneme embeddings, which are combined with prosodic features with the Phoneme–Prosody Cross-Attention Fusion (PP-CAF) layer. The fused representations are then scored by a multi-task Transformer scoring model in order to determine the accuracy of pronunciation, quality of prosody, and intelligibility. Lastly, adaptive curriculum feedback is created on

learners and the output is brought together through the evaluation and reporting process, which allows personalized and interpretable results in performance.

A. Data Collection

The Speech Accent Archive dataset includes 2,140 speech samples from individuals representing 214 native languages in 177 countries, each of whom reads the same standardized English passage. It contains demographic details, enabling analysis of various factors such as age, gender, original language and speakers. The core of the dataset contains authentic audio recordings (.MP3) of indigenous and non-indigenous English speakers, which support accent, phonetic variations and controlled comparison of processed patterns. This resource is invaluable for ASR training, phonetic research, and prosecution analysis; the convenience of development of pronunciation recognition models; accent assessment tools; and mitigation in speech AI systems [25]. It applies the Speech Accent Archive (SAA) as the main training and test data. Audio samples are all phoneme aligned, normalized and resampled. To test cross-lingual robustness of PP-CAF and AG-CLSSL scoring system, additional samples of multilingual learner speech were also added.

B. Data Pre-Processing

The step of preprocessing guarantees the transformation of raw audio recordings into standardized and noise-free raw materials that can be utilized in the extraction of features and model training. Because the Speech Accent Archive dataset consists of recording of different speakers under variable recording conditions, preprocessing is essential to guarantee consistency across samples and to make sure that the resulting feature extraction captures pronunciation and prosody patterns instead of recording artifacts.

1) Signal cleaning and normalization: All audio recordings are resampled to a uniform frequency of 16 kHz to standardize the temporal resolution. Background noise is reduced using spectral subtraction, while amplitude normalization ensures consistent loudness across samples. This procedure reduces the variability caused by different recording devices and environments, allowing the model to focus solely on learner-specific speech characteristics.

2) Aegmentation and alignment: The segmentation and alignment are applied to map learner speech at the phoneme and word levels. Forced alignment techniques are used to synchronize the learner's utterance with a native benchmark transcription. This alignment enables the identification of phoneme-level insertions, deletions, and substitutions, which are crucial for detecting mispronunciations. The alignment process is represented in Eq. (1),

$$A(t) = Align(S(t), R(t)) \tag{1}$$

where, $A(t)$ denotes the alignment mapping at time t, $S(t)$ is the learner's speech signal, and $R(t)$ is the reference transcription. The output is a time-aligned phoneme sequence that highlights deviations between learner and native speech.

3) Prosody feature preparation: In addition to segmental accuracy, suprasegmental features such as stress, rhythm, and intonation play an important role in overall fluency. Therefore, prosody feature preparation is performed by extracting pitch (fundamental frequency, F_0), duration, and intensity measures. These parameters capture temporal and melodic variations in speech. The prosody vector for each utterance is defined in Eq. (2):

$$P(u) = [\mathrm{F_0(u)}, \mathrm{D(u)}, \mathrm{I(u)}] \tag{2}$$

where, $P(u)$ represents the prosody feature vector for utterance u, $\mathrm{F_0(u)}$ is the pitch contour, $\mathrm{D(u)}$ is the duration of phonemes, and $\mathrm{I(u)}$ is the intensity or loudness profile. These features form the foundation for analyzing suprasegmental errors such as misplaced stress or unnatural intonation patterns.

C. Feature Extraction

The XLS-R (wav2vec2-large-xlsr-53) serves as the backbone feature extractor. XLS-R is a self-supervised, cross-lingual model that was trained on millions of hours of multilingual speech. First, the resample raw learner audio to 16 kHz mono and pass it through the XLS-R model to obtain high-dimensional contextual embeddings that capture both fine-grained acoustic–phonetic detail and longer-range linguistic context. Then, identify phoneme boundaries using the Montreal Forced Aligner and aggregate frame level embeddings from XLS-R within each phoneme boundary to form more robust phoneme-level vectors. To stabilize the representation, average the hidden states of the last four transformer layers of XLS-R to balance lower-level acoustic cues with higher-level contextual shifting information. The output of XLS-R is represented as H in Eq. (3).

$$H = XLSR(x) = [h_1, h_2, \dots, h_T],\ h_t \epsilon \mathbb{R}^d \tag{3}$$

Where, x is the input speech waveform resampled to 16kHz mono; $h_t \epsilon \mathbb{R}^d$ is the embedding vector at time frame t, where $d = 1024$ is the hidden dimension of XLS-R and T is the number of time frames (sub word units) produced by XLS-R for the utterance. The Phoneme-level pooling is denoted in (4).

$$h_p = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} h_t \tag{4}$$

Here, h_p is Phoneme-level embedding for phoneme p, computed by pooling frame embeddings within its boundaries; t_s and t_e are the start and end frame indices of a given phoneme, obtained from forced alignment. The Projected phoneme embedding in a lower-dimensional shared space denoted as $\tilde{h}_p$ in Eq. (5) and the Projected prosody embedding in the same shared space is represented as $\tilde{f}_p$ in Eq. (6):

$$\tilde{h}_p = W_p h_p + b_p,\ \ \tilde{h}_p \epsilon \mathbb{R}^d \tag{5}$$

$$\tilde{f}_p = W_f h_f + b_f,\ \ \tilde{f}_p \epsilon \mathbb{R}^d \tag{6}$$

where, W_p is the learnable projection parameters for transforming phoneme embeddings, and W_f is the earnable projection parameters for prosody features.

Fig. 2 illustrating the self-supervised pre-training stage and supervised fine-tuning stage. In the self-supervised stage, raw audio is processed by a convolutional feature encoder, masked prediction, and a Transformer network, which produces

contextualized speech embeddings, with quantization as a technique for improving stability of the representation learning process. The supervised fine-tuning stage adopts these pre-trained embeddings with learner specific data, enabling the transfer of knowledge to distinguish pronunciation and prosody-based features. The outputs are incorporated into a personalized feedback module, which supports adaptive and targeted feedback for learners to improve the segmental and suprasegmental components of speech. After obtain the phoneme embeddings, map the embeddings into 512-dimensional space to reduce computational efficiency, then feed the representations into the intended PP-CAF Layer, where embeddings are dynamically aligned with prosodic features (pitch, energy, and duration). By grounding phoneme representations in acoustic detail and suprasegmental cues, XLS-R serves as powerful backbone that provides the proposed framework to evaluate pronunciation and prosody simultaneously with improved representation of the learner and hence improved performance over traditional concatenating features or one stream processing.

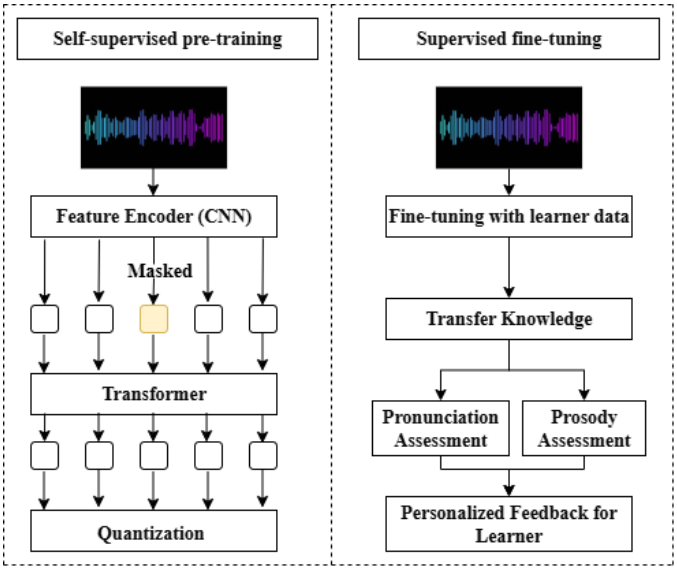

Fig. 2. Architecture of the proposed framework.

D. Phoneme–Prosody Cross-Attention Fusion

The analysis presents the PP-CAF Layer that directly combines segmental and suprasegmental cues via cross-attention as opposed to concatenation. Embeddings of phonemes, pooled at XLS-R outputs across the boundaries between phonemes, are attended to by projected prosodic features, and yield the resultant enriched vectors, which integrate the phonemic accurateness with the prosodic appropriateness. These combined representations are then processed through a multi-task Transformer scorer to perform consonant, vowel, and consonant together predicting pronunciation, prosody, and intelligibility. The, PP-CAF allows articulation and prosody to interact modelling only those cases where phonemes themselves are correct but are stressed or intonated in the wrong way, providing a more holistic and context-relevant view of the speech of a learner than has been previously possible. The PP-CAF layer implements scaled dot-product cross-attention is expressed in Eq. (7).

$$A_p = softmax\left(\frac{\tilde{h}_p W_q \left(\tilde{f}_p W_k\right)^T}{\sqrt{d_k}}\right) \quad (7)$$

where, W_q, W_k $\epsilon \mathbb{R}^{d' \times d'}$ are learnable projection matrices, and d_k is the attention dimension. The enriched phoneme-prosody vectors are passed into the multi-task Transformer scorer.

E. Scoring Model with Multi-task Transformer

The integrated representations emitted from the PP-CAF layer are funneled into a multi-task Transformer-based scoring model that evaluates pronunciation accuracy, prosody quality, and overall intelligibility concurrently. Each phoneme–prosody vector is first input into the positional encoding layer, then into a 6-layer Transformer encoder that captures dependencies across phonemes and syllables in an utterance. This architecture allows the model to consider local errors (e.g., phoneme substations) and global structures (e.g., stress and rhythm across a phrase) at the same time. Two multi-task output heads are then implemented on top of the common Transformer encoder. The first output head predicts whether each phoneme was pronounced correctly, and this classification is compiled as the score for pronunciation accuracy. The second output head scores prosody regression, predicting the stress and rhythm alignments with reference scores to compute a prosody quality score. The third output head use regression to obtain an intelligibility assignable with the 1-5 scales of human experts. To train / tune the model, collect and apply a weighted multi-task loss function to the three tasks. It is represented in Eq. (8):

$$L = \alpha L_{pron} + \beta L_{prosody} + \gamma L_{intell} \quad (8)$$

where, L_{pron} is a cross-entropy loss for phoneme correctness classification, $L_{prosody}$ is a mean squared error loss for prosody scoring, and L_{intell} is an L1 regression loss for intelligibility. The weights α, β and γ are tuned empirically on the validation set to ensure balanced learning across tasks.

Algorithm 1: Attention-guided Cross-Lingual Self-Supervised Learning

Input:
- Speech utterance x (16 kHz mono)
- Reference transcript T
- Forced alignment boundaries B = {b1, b2, …, bn}

Preprocess input audio (resample, trim silence, normalize).

Extract contextual embeddings H = XLS-R(x).

For each phoneme boundary (ts, te) in B:
- Pool XLS-R embeddings → phoneme vector hp

Extract prosody features for each phoneme:
- Pitch (YAAPT), Energy (RMS), Duration (MFA)
- Project features into embedding space fp

Apply PP-CAF:
- Fused representation up = Attention (hp, fp)

Feed fused representations U = {u1, u2, …, un} into Transformer encoder.

Apply task-specific output heads:
- Pronunciation head → accuracy score
- Prosody head → stress/rhythm score

Intelligibility head → 1–5 rating
Compute total loss:
$L = \alpha L_{pron} + \beta L_{prosody} + \gamma L_{intell}$
Output:
Pronunciation accuracy score
Prosody quality score
Intelligibility score
Adaptive feedback for learner

Algorithm 1 outlines the Attention-guided Cross Lingual Self Supervised Learning framework presented in this study, which combines pronunciation and prosody assessment in a single learning pipeline. The system initially preprocesses the audio of the learner and derives phonemically embedded XLS-R as well as prosodic (pitch, energy, duration, etc.) features. The PP-CAF layer combines these two streams into a single stream generating better representations that are more segmental and suprasegmental. Fused vectors are then fed into a multi-purpose Transformer in order to concurrently predict pronunciation accuracy, prosody quality, and intelligibility. The weighted loss function leads to model training and in the course of the study, adaptive curriculum scheduler offered customized multimodal feedback as a support to robust assessment as well as to learner advancement. The sensitivity analysis involved a parameter sensitivity analysis in the learning rate, weights of the PP-CAF fusion, and multi-task loss coefficients. Findings depict that the fusion weight (α) is a powerful predictor of prosody accuracy and loss weight (λ) is a powerful predictor of stability in scoring pronunciation. Any slight differences in learning rate caused very little performance variation.

The suggested procedure combines both segmental and suprasegmental speech analysis in one learning path, which is a novel concept in the pronunciation evaluation. The homophonic characteristics of the traditional tools pay attention to phoneme-level accuracy most, with prosody being a secondary characteristic, in some cases limited to length. The interaction of phoneme embeddings with prosodic cues, such as pitch, energy and duration, is possible with the PP-CAF layer allowing a more thorough framework to be presented to cover the instances of phonemes being pronounced correctly but the prosody being unnatural. The method will make use of XLS-R self-supervised embeddings, which offer multilingual capabilities, encode finer-grained phonemic and contextual features and are resistant to learners with varying language backgrounds. A multi-task Transformer scoring model is used to process these enriched representations and makes joint predictions of the accuracy of the pronunciation, the quality of the prosody and the intelligibility. The methodology, which simulates the interaction between articulation and suprasegmental features, provides a holistic and context-sensitive measurement of the spoken language which is superior to other conventional assessment which considers segmental and prosodic information independently.

IV. Results and Discussion

The suggested Attention-guided Cross-Lingual Self-Supervised Learning model significantly increased a performance of learners in spoken English at all levels of proficiency. The framework generated more detailed speech representations, with the addition of phoneme embeddings and prosodic cues via the PP-CAF layer, which allowed learners to perform phonemic articulation more accurately and had a manufacturing speech that was more understandable and fluent and easier to comprehend. Prosodic aspects were strengthened, and there was the increased natural stress placement, less stumbling rhythm, and regular intonation patterns of the entire speech. Adaptive feedback of the framework offered a real time, customized feedback and gave learners the opportunity to rectify articulation and prosodic mistakes in real time when conversing. The feedback was provided in a multimodal way - visual pitch contours, auditory examples, and textual hints, which enhanced the activity of learners and the intensity of their practice. The participants expressed increased confidence, less anxiety, and increased motivation to engage in future verbal communication activities. Altogether, the framework promoted a comprehensive enhancement in the areas of pronunciation, fluency and prosody, as well as the establishment of a conducive, interactive and encouraging learning atmosphere.

Table I shows the Simulation parameters used in the application of the proposed framework. The table presents a summary of the experimental design applied to guarantee similar preprocessing, feature extraction, model setup, and training procedures. These parameters were set to guarantee good reproducibility and computational stability and reliable results for assessing performance of the proposed framework for all case studies.

TABLE I. Simulation Parameter Table

Parameter	Value
Sampling rate	16 kHz
Frame length (for prosody / F0)	25 ms
Frame hop	10 ms
Alignment granularity	phoneme-level
Pitch frame step	10 ms
XLS-R checkpoint	wav2vec2-large-xlsr-53
Projected phoneme dim (d′)	512
Prosody embedding dim	512
Encoder layers	6
Model dim	512
Feed-forward dim	2048
Attention heads	8
Positional encoding	sinusoidal or learned
Pronunciation head	phoneme-level classification
Batch size	16 (adjust to GPU mem)
Epochs (head training)	20–30
Warmup steps	2,000
Weight decay	0.01
Gradient clipping	max-norm = 1.0
Dropout	0.1
Random seeds	{42, 123, 2024}
GPU	NVIDIA V100 / A100 recommended
Validation split	speaker-independent val set (10–15%)

A. Experimental Outcome

The Experimental outcome of the proposed framework yielded significant improvement in learner speech production, including more precise phoneme production, more fluid rhythm, and more natural intonation. Learners were able to adapt quickly to the new system, participated actively with the multimodal feedback, and increased their self-awareness of errors as they corrected them during practice. The adaptive curriculum models provided a personalized practice journey for each participant that allowed for more balanced training progress. Participants also expressed increases in confidence, motivation, and willingness to communicate in spoken English.

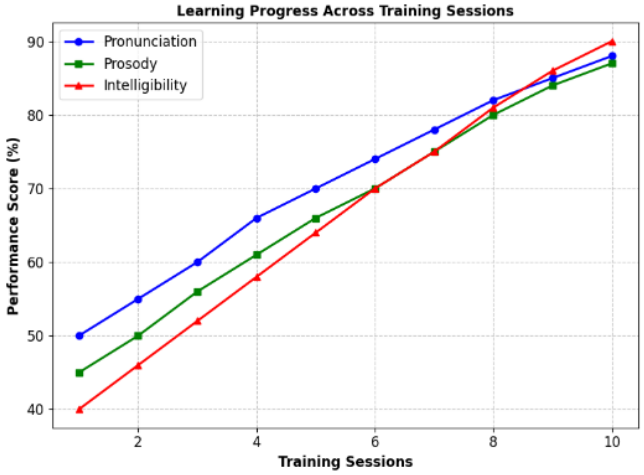

Fig. 3. Learning progress across training sessions.

Fig. 3 shows continuous and marked improvements in pronunciation accuracy, prosody quality and overall intelligibility across the ten sessions. Each of the three dimensions showed a consistent and positive growth trajectory, indicating that the system supported improved performance on multiple aspects of spoken English for individual learners simultaneously.

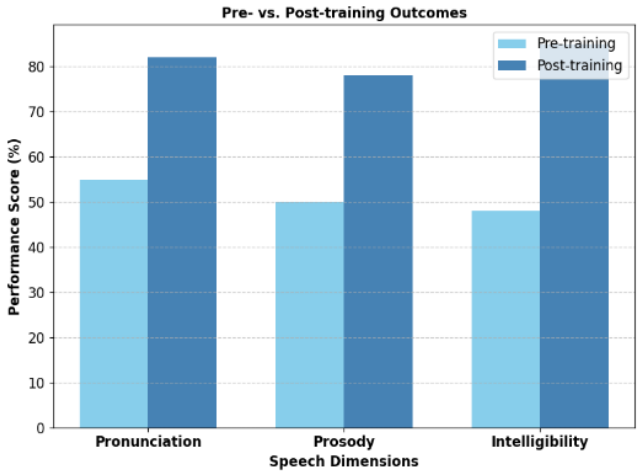

Fig. 4. Pre vs. Post training outcomes.

Fig. 4 depicts clear differences across three dimensions – pronunciation, prosody, and intelligibility – after training, with the differences persisting and consistent following the training period. The increased performance across the three dimensions indicates that the framework is effective in developing and enhancing both segmental and suprasegmental features of speech.

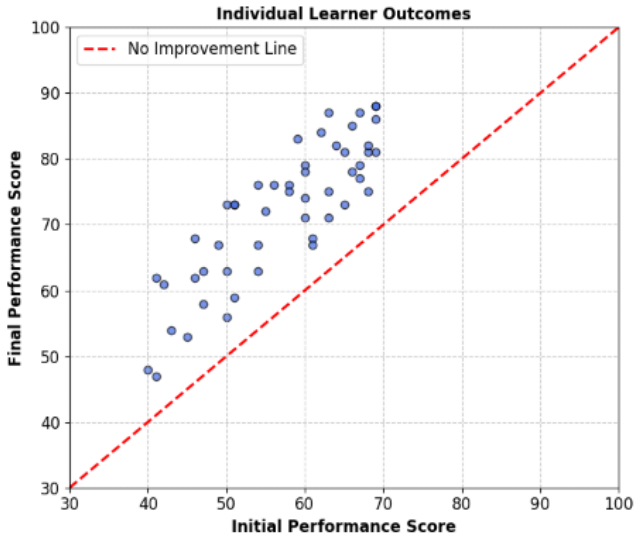

Fig. 5. Individual learner outcome.

Fig. 5 illustrates the individual learner outcomes before and after training with the proposed framework. Each point represents a learner's original and final score for performance. Most points are above the diagonal reference line indicating that the vast majority of learners improved their pronunciation, prosody and overall intelligibility, while using the proposed framework.

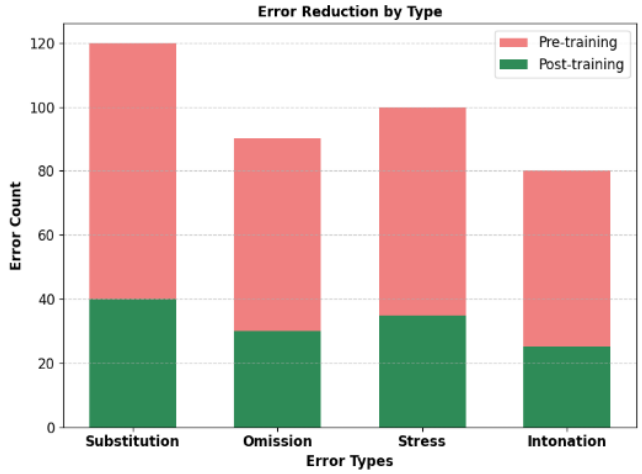

Fig. 6. Error reduction by type.

Fig. 6 demonstrating reduction of errors by type after training with the Attention-guided Cross Lingual Self Supervised Learning framework. The number of errors demonstrated a decrease in the following categories: substitution, omission, stress, and intonation, illustrating an overall improved balance of precision in articulation and prosody.

Fig. 7 depicting learner development at levels of proficiency over time in training using the proposed framework. Each group,

beginner, intermediate, and advanced, consistently showcased growth; beginners experienced the most acceleration in productivity, the intermediate learners displayed steady growth, and the advanced group's development changed in smaller increments. This growth trajectory reinforces the proposed framework's effectiveness in meeting learners' needs at different levels of language proficiency.

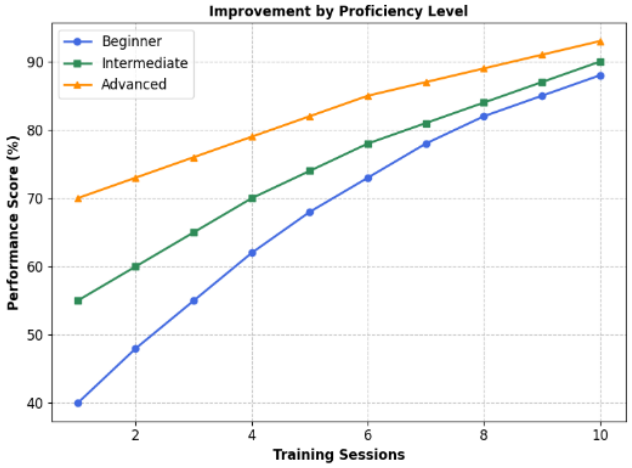

Fig. 7. Improvement by proficiency level.

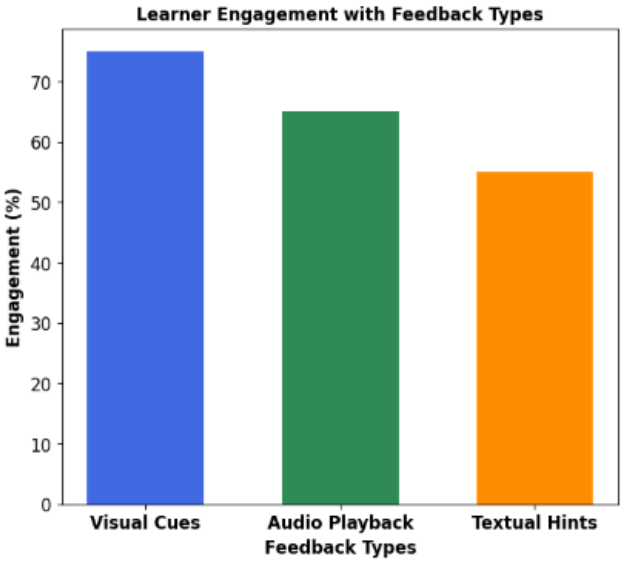

Fig. 8. Learner engagement with feedback types.

Fig. 8 shows learner engagement with each type of feedback in the Attention-guided Cross Lingual Self Supervised Learning framework. Visual cues were the most utilized resources, followed by audio playback and textual hints. This demonstrates the benefits associated with multimodal feedback, as learners clearly favored visual and auditory assistance in learning about pronunciation and prosody.

Fig. 9 shows a two-dimensional cluster plot of learner outcomes at the end of the training period with the Attention-guided Cross Lingual Self Supervised Learning framework. Each point represents a learner and their location within the plot is based on their articulation and prosody scores. The color-coded clusters illustrate naturally occurring groupings in performance and provide an understanding of how the proposed framework enabled unique paths toward improvement in speech proficiency.

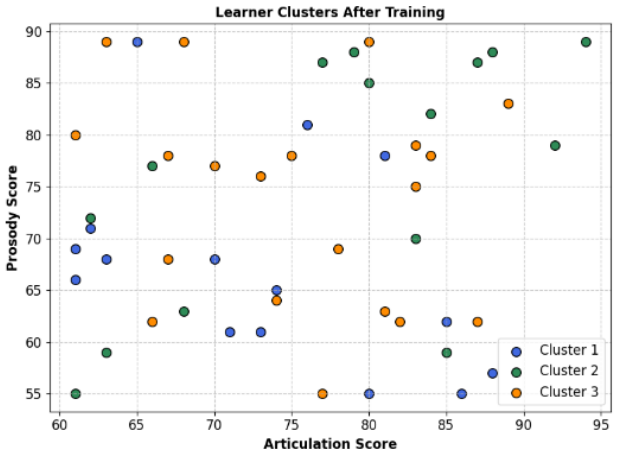

Fig. 9. Learner cluster after training.

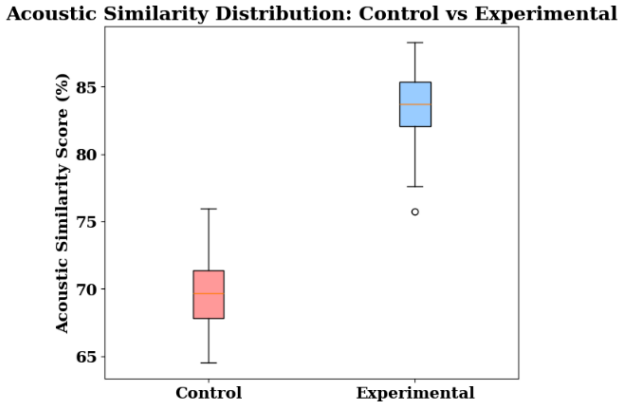

Fig. 10. Acoustic similarity distribution: Control vs. Experimental.

Fig. 10. compares the acoustic similarity scores of controls and experimental groups post-training. The experimental group achieved significantly higher median and tighter distribution around 83.6%, while the control group centered near 70.4% with broader variability. This shows that the Attention-guided Cross Lingual Self Supervised Learning training provides more stable and natural pronunciation patterns than traditional training.

B. Performance Evaluation

Attentionguided Crosslingual Selfsupervised Learning (ACLSL) model was also tested both by the system generated outputs and the judgment of the expert, it was shown that the model can integrate both the phoneme level articulation with the suprasegmental prosody to generate the complete representation of the speech. The analyses of learners recorded also showed significant growth in the accuracy of phonemes, rhythmic fluency, and stress realization as well as the intonation patterns before and after training. Objective system improvements were supported by expert rater reports that the post-training speech was always more natural, intelligible, and fluent. Students found out that the adaptive feedback system helped them to practice more effectively as they could self-correct articulation and prosodic mistakes in real-time. The visual contours of pitch, the

auditory examples and the textual cues were all combined as the multimodal feedback and this strengthened pattern recognition and encouraged correct production. Computational analyses showed that the framework was stable in terms of performance regardless of the level of learner proficiency and language backgrounds and ensured the reliability and robustness of the system when used alone. The multi-task Transformer scoring model has been able to learn both segmental and suprasegmental of some speech at once, and hence no individual dimension was improved while other dimensions were neglected. Therefore, the ACLS model facilitated holistic improvement of pronunciation, prosody, intelligibility and general production of the spoken language offering learners, who had varying learning needs, a scaled, context sensitive and adaptive learning platform.

1) Phoneme error rate: Another significant measure to be used to assess the pronunciation accuracy at phoneme level is PER. It measures the proportion of phoneme errors, i.e. substitutions, additions, and deletions, relative to a native pronunciation point of reference, which is provided in Eq. (9).

$$\mathrm{PER} = \frac{S+D+I}{N} \times 100 \quad (9)$$

Where *S* denotes the substitutions, *D* denotes the deletions, *I* denotes the insertions, and *A* denotes the total phonemes. A decrease in PER implies a reduction in the number of mistakes in articulation, and accuracy in phonemes.

2) Word error rate: WER is concerned with the entire-word pronunciation accuracy and phoneme sequence errors make up words, whereas PER is concerned with phoneme-level accuracy. In this manner, it will be able to determine the extent to which mistakes in phonemes are affecting word intelligibility assembled in a conversational context, that is provided in Eq. (10):

$$\mathrm{WER} = \frac{S+D+I}{W} \times 100 \quad (10)$$

where, *W* is the number of words. Moreover, a decrease in the values of WER signifies some positive changes in the intelligibility of the overall spoken utterance.

3) Acoustic similarity score: The Acoustic Similarity Score is a rule of how intake pronunciation produced by a learner is similar to that of native pronunciation patterns using embeddings produced through deep learning. It is expressed in Eq. (11):

$$\mathrm{ASS} = \frac{A_L \cdot A_N}{\|A_L\| \|A_N\|} \times 100 \quad (11)$$

where, A_L and A_N represent learner and native embeddings respectively. Moreover, higher scores demonstrate that there is a propensity to more naturalistic production.

4) Pitch alignment: One is a parameter Pitch Alignment. The larger the difference between the reference pitch, the more one can notice the occurrence of intonation errors: monotonicity of the delivery or inappropriate placement of the rising and falling intonation marks that may very well disrupt the clarity and expressiveness of the speech severely in Eq. (12).

$$\mathrm{PAS} = 1 - \frac{\|P_L - P_N\|}{\|P_N\|} \times 100 \quad (12)$$

where, P_L and P_N represent earner and native pitch trajectories. There are more accurate patterns of intonation as well as higher scores.

5) Duration consistency index: Duration Consistency Index determines timing and rhythm of syllables and words in the speech of learners in comparison with native speakers.Natural speech rhymes in its rhythmic patterns and observes variations in relative syllable length, word duration, and pause. EFL learners often encounter pacing problems and speak too fast or insert pauses where inappropriate within words or phrases. The metric analyzes the speech waveform-performed segmentation by temporal alignment models and assesses how much the native timing patterns are deviated from. A very high consistency score indicates smooth and rhythmic speech, and a very low score indicates problems in fluency and speech timing, that is represented in Eq. (13).

$$\mathrm{DCI} = 1 - \frac{\sum |T_L - T_N|}{\sum T_N} \times 100 \quad (13)$$

Here T_L and T_N are the learner and native durations. Increased scores reflect that timing smoother and improved flow of speech.

6) Stress placement accuracy: In the case of English, Stress Placement Accuracy is significant to the intelligibility as unlikely stress may cause misunderstanding. The score of high accuracy is decipherable as a natural sound; a failure in the correct execution of stress might cause prospective miscommunication, otherwise, not a very robot-like speech is delivered in Eq. (14):

$$\mathrm{SPA} = \frac{C_S}{T_S} \times 100 \quad (14)$$

where, C_S is the properly stressed syllabus and T_S is the total syllabus. Increased scores are more indicative of natural and intelligible stress patterns.

7) Intensity deviation: Variation in intensity is one of the effective prosodic features which practically means expressiveness and articulateness of speech. The above described measurements have the benefit that they allow identifying where stress is misplaced and the irregularities of the speech volume are in the effort of the learner towards more natural and expressive speech patterns are depicted in Eq. (15).

$$\mathrm{ID} = \frac{|I_L - I_N|}{I_N} \times 100 \quad (15)$$

Here, I_L and I_N are the learner and native intensity values respectively. Lower scores likely indicate better similarity to natural expressiveness.

C. Session Frequency and Duration

The intensity and duration of learners are important behavioral measures of the involvement of learners in the Attention-guided Cross Lingual Self Supervised Learning based pronunciation training. With the help of such a correlation, it is

possible to inform systems about the most effective patterns of the engagement and individualize the training strategies, optimizing the fluency and pronunciation improvements after training in EFL learners.

D. Retry and Correction Behavior

The Attention-guided Cross Lingual Self Supervised Learning based pronunciation training is significantly dependent on repetition and correction as measures of self-regulated learning. The effort of every learner is recorded; the difference being recognized between the willingness to complete the task and sincere efforts to become better with pronunciation. The retry-correction tracking assists in customizing procedures of learning, ensuring interest, and preserving durability of fluency and accuracy skills.

E. Dropout Rate and Fatigue Detection

The vital elements covered in the maintenance of Attention-guided Cross Lingual Self Supervised Learning stimulated pronunciation instruction were dropout and fatigue. The introduced AI-based technological solution stimulates long-term motivation and, thus, allows the learners to continue moving in the right direction and ultimately master pronunciation.

F. Self-Assessment Surveys

Although self-assessment surveys are subjective in nature, they provide useful information as to the confidence of the learners, their perceived progress and their satisfaction with Attention-guided Transformer Cross Lingual Self Supervised Learning based pronunciation training. The system is able to personalize effective pronunciation instruction that provides a comprehensive learner profile by combining insights of self-assessment with the AI-driven analytics, which ensures that the cognitive and emotional learning needs of the learner are met.

G. Expert Evaluations

There are two ideas that are conveyed with this text. One of them is that expert judgments can have two functions other than being validation on AI-based pronunciation assessment. The interaction between human expert judges and AI assessment surrogates breeds a balanced scoring system whereby the system promotes performance of objective judgment on the quality of pronunciation besides integrating the qualitative to the human input.

H. Error Clustering Insights

The clustering of phoneme errors that are NLP-based can offer a deeper understanding of the pronunciation challenge peculiar to the language as a whole as a cluster of similar errors shared between learners in regard to their influence of the first language, phoneme substitutions, and articulation patterns. Task specific feedback would be provided to the learners, covering corrections regarding the most common pronunciation glitches of the learners also being linguistically relevant, thus making Attention-guided Transformer Cross Lingual Self Supervised Learning training even more fine-tuning.

I. Comparative Analysis

The study assessed the effectiveness of Attention-guided Transformer Cross Lingual Self Supervised Learning based pronunciation training in comparison to traditional methods, such as teacher-led phonetics, textbook drills, and classroom repetition exercises. Additionally, learner self-assessments and engagement surveys were analysed to capture the motivational and confidence-boosting effects of Attention-guided Cross Lingual Self Supervised Learning based training.

TABLE II. COMPARATIVE ANALYSIS

Metric	Pre-Training (%)	Post-Training (%)	Improvement (%)
PER	28.4%	22.7%	55.9%
WER	21.5%	18.2%	51.1%
Pitch Alignment Score	61.4%	68.7%	35.2%
Duration Consistency Index	59.3%	65.1%	35.3%
Stress Placement Accuracy	57.1%	63.4%	43.7%

Table II shows the pre-and post-training evaluation of learners using the Attention-guided Transformer Cross Lingual Self Supervised Learning framework. The results indicate consistent improvement across phoneme-level, word-level, and prosody-related measures, including reduced phoneme and word errors, stronger prosodic alignment, and improved expressiveness. Intensity deviation improved significantly, more natural delivery.

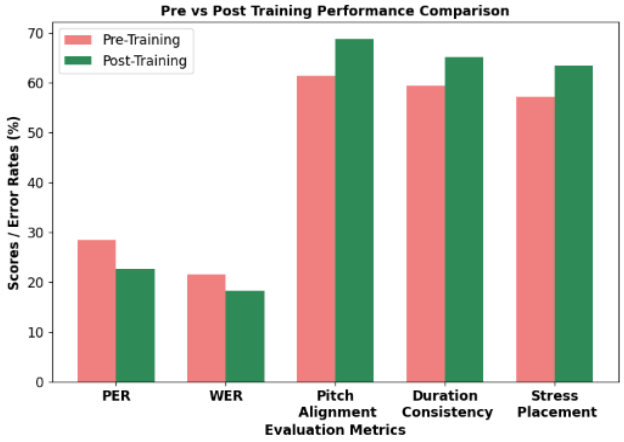

Fig. 11. Comparative analysis.

Fig. 11 shows the comparison of pre-training and post-training performance across five evaluation metrics. The Attention-guided Transformer Cross Lingual Self Supervised Learning framework resulted in consistent improvements by decreasing error rates for the PER and WER measures, and increasing score values for prosodic measures of Pitch Alignment, Duration Consistency, and Stress Placement Accuracy. These results underscore the ability of the proposed model to improve pronunciation accuracy and prosodic quality.

J. Discussions

The Attention-Guided Cross-Lingual Self-Supervised Learning (AG-CLSSL) system can be seen as a valuable step forward toward making EFL learners speak better in terms of pronunciation and prosody. AG-CLSSL achieves accuracy in both segmental and suprasegmental dynamics by incorporating

the XLS-R phoneme-level embeddings using new suprasegmental cues, pitch, duration, and energy, which combine into single, context-sensitive speech representations, the Phoneme-Prosody Cross-Attention Fusion (PP-CAF) layer. The experimental outcomes show a significant decrease in Phoneme Error rate (PER) and Word Error rate (WER), significant changes in stress placement and prosody alignment, which need to be considered as the problem that is critical to address the issue of articulation and suprasegmental aspects, which were traditionally considered separately. Pedagogically, the framework allows flexible, learner-based training through a multi-task Transformer, and provides tailored feedback, which is aligned to the level of proficiency without causing cognitive overload among learners. The integrative method increases objective acoustic accuracy and increases the perceived naturalness, intelligibility and confidence of the learners. What is more, the attention guided architecture can enable interpretability whereby educators and learners can identify areas to improve. AG-CLSSL though mainly tested in controlled contexts has potential in real-time classroom applications and incorporation of more speech parameters and the development of CAPT is through scalable, pedagogically robust and interpretable cross-linguistic self-supervised modeling.

V. Conclusion and Future Works

The research presented a complex and attention-based model AG-CLSSL, which aimed at improving pronunciation and prosody in EFL learners simultaneously. Through XLS-R embeddings in combination with prosodic cues, which are provided with the use of the proposed PP-CAF layer, the system generates a single, context-sensitive speech representation, allowing to score phoneme quality, prosodic alignment, and general intelligibility with accuracy when using multi-task scoring. As a result of the experiment, significant improvements in the articulations, stress placement, rhythm, pitch contour, and naturalness are proven, which validates the effectiveness of the framework. The patterns of visualizable attention also lend credence to transparent center learner feedback, making the model very beneficial in the application of CAPT. In general, AG-CLSSL offers a scaffoldable, pedagogically significant strategy that results in the development of technical performance and practical learning outcomes.

Future studies will center their attention on measuring the scalability and the generalizability of AG-CLSSL in large groups of learners and in various learning contexts. The addition of the prosodic feature set which covers pause patterns, speech rate, and discourse-level rhythm can also contribute to the improvement of fluency and expressiveness assessment. The inclusion of real-time, interactive learning resources, including adaptive or even gamified interfaces, might reinforce the motivation and the engagement of the learners. Also, cross-linguistic research will be done to determine the applicability of the framework in other language and dialects. In general, these directions will bring AG-CLSSL into a more flexible, robust and pedagogically effective system to next-generation intelligent language learning systems.

References

[1] "(PDF) Why is Pronunciation So Difficult to Learn?" Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/265821016_Why_is_Pronunciation_So_Difficult_to_Learn

[2] "(PDF) The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis," ResearchGate, Dec. 2024, doi: 10.1017/S0958344023000113.

[3] "(PDF) Pronunciation Assessment: Traditional vs Modern Modes," ResearchGate, Oct. 2024, doi: 10.56916/jesi.v1i1.530.

[4] "Using Technology for Pronunciation Teaching, Learning, and Assessment: Contemporary Perspectives | Request PDF." Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/327536418_Using_Technology_for_Pronunciation_Teaching_Learning_and_Assessment_Contemporary_Perspectives

[5] "(PDF) A study of anxiety experienced by EFL students in speaking performance," ResearchGate, Oct. 2024, doi: 10.24815/siele.v7i2.16768.

[6] "(PDF) A Survey of Automatic Speech Recognition for Dysarthric Speech." Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/374770801_A_Survey_of_Automatic_Speech_Recognition_for_Dysarthric_Speech

[7] "(PDF) The Role of ASR Training in EFL Pronunciation Improvement: An In-depth Look at the Impact of Treatment Length and Guided Practice on Specific Pronunciation Points," ResearchGate. Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/362684159_The_Role_of_ASR_Training_in_EFL_Pronunciation_Improvement_An_In-depth_Look_at_the_Impact_of_Treatment_Length_and_Guided_Practice_on_Specific_Pronunciation_Points

[8] "(PDF) Perception of Prosodic Speech Features: Final Intonation and Word Stress for EFL Learners," ResearchGate, Dec. 2024, doi: 10.37999/udekad.1449612.

[9] "(PDF) Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," in ResearchGate, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/250306074_Automatic_Error_Detection_in_Pronunciation_Training_Where_we_are_and_where_we_need_to_go

[10] "(PDF) Dynamic difficulty adjustment using deep reinforcement learning: A review," ResearchGate, Oct. 2024, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/383664462_Dynamic_difficulty_adjustment_using_deep_reinforcement_learning_A_review

[11] "(PDF) Correcting Mispronunciations in Speech using Spectrogram Inpainting," in ResearchGate, doi: 10.21437/Interspeech.2022-615.

[12] "Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults | Request PDF," ResearchGate, Dec. 2024, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/346641196_Effects_of_an_automatic_speech_recognition_system_with_peer_feedback_on_pronunciation_instruction_for_adults

[13] "(PDF) The Implementation of Automated Speech Recognition (ASR) in ELT Classroom: A Systematic Literature Review from 2012-2023," ResearchGate, Dec. 2024, doi: 10.29408/veles.v7i3.23978.

[14] "Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: a mixed-methods study," ResearchGate, Oct. 2024, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/353474805_Mobile-assisted_pronunciation_learning_with_feedback_from_peers_andor_automatic_speech_recognition_a_mixed-methods_study

[15] M. Orosoo et al., "Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education," Alex. Eng. J., vol. 111, pp. 21–32, Jan. 2025, doi: 10.1016/j.aej.2024.10.065.

[16] "Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning | Request PDF," in ResearchGate, doi: 10.21437/Interspeech.2022-10245.

[17] "(PDF) Intelligibility and the Listener: The Role of Lexical Stress," ResearchGate, Oct. 2024, doi: 10.2307/3588487.

[18] "Visual Prosody and Speech Intelligibility Head Movement Improves Auditory Speech Perception | Request PDF," ResearchGate, Oct. 2024, Accessed: Mar. 07, 2025. [Online]. Available:

https://www.researchgate.net/publication/8341805_Visual_Prosody_and_Speech_Intelligibility_Head_Movement_Improves_Auditory_Speech_Perception

[19] "(PDF) Automatic Speech Recognition Using Limited Vocabulary: A Survey," ResearchGate, Dec. 2024, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/362241766_Automatic_Speech_Recognition_Using_Limited_Vocabulary_A_Survey

[20] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised Automatic Speech Recognition: A review," Speech Commun., vol. 139, pp. 76–91, Apr. 2022, doi: 10.1016/j.specom.2022.02.005.

[21] "The effects of automatic speech recognition quality on human transcription latency," in ResearchGate, doi: 10.1145/2899475.2899478.

[22] "(PDF) A Study on the Situation of Pronunciation Instruction in ESL/EFL Classrooms," ResearchGate, Feb. 2025, doi: 10.5296/jse.v1i1.924.

[23] "Title: Phonological Differences between Japanese and English: Several Potentially Problematic Areas of Pronunciation for Japanese ESL/EFL Learners," ResearchGate, Oct. 2024, Accessed: Mar. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/239806899_Title_Phonological_Differences_between_Japanese_and_English_Several_Potentially_Problematic_Areas_of_Pronunciation_for_Japanese_ESLEFL_Learners

[24] "(PDF) The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation," ResearchGate, Oct. 2024, doi: 10.3389/fpsyg.2023.1210187.

[25] "Speech Accent Archive." Accessed: Mar. 07, 2025. [Online]. Available: https://www.kaggle.com/datasets/rtatman/speech-accent-archive