An Articulatory-Aware CNN-BiGRU-Attention Framework for Explainable Phoneme-Level Pronunciation Assessment in ESL Speech

Dr P. Bindhu[1], Jasgurpreet Singh Chohan[2], Dr. M. Durairaj[3], Dr. Megha Sawangikar[4],
Dr. N. Neelima[5], Elangovan Muniyandy[6], Dr. G. Sanjiv Rao[7], Loay F. Hussien[8]
Assistant Professor of English, Department of Humanities and Science, Rajalakshmi Institute of Technology, Chennai, India [1]
Marwadi University Research Center-Department of Mechanical Engineering-Faculty of Engineering & Technology, Marwadi University, Rajkot, Gujarat, India[2]
Department of English, Panimalar Engineering College, Chennai, India[3]
Assistant Professor, Department of Applied Chemistry, Yeshwantrao Chavan College of Engineering, Wanadongri, Nagpur, Maharashtra, India[4]
Associate Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, 522302, Andhra Pradesh, India[5]
Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105[6]
Professor, Dept. of AIML, Aditya University- Suramplem, AP, India[7]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia[8]

***Abstract*—Proper pronunciation at the phoneme level has been known to be one of the most enduring problems affecting the Second Language learners of the English language (ESL) since the slight pronunciatory variations in the learned language may greatly influence its communicative power and the level of intelligibility. The existing methods of pronunciation evaluation, which are mostly made using automatic speech recognition (ASR), place their results at the word level or the sentence level and offer generic numerical scores with little linguistic meaning, which is not effective in assessing accented speech and subsequent correction. To overcome these shortcomings, the paper introduces an articulatory-conscious recognition model of phonemes that provides fine-grained and interpretable feedback to enhance ESL pronunciation. The novelty of the work is in the combination of a hybrid CNN-BiGRU-Attention architecture and an Articulatory Error Mapping Engine, which symbolically transforms phoneme-level articulation errors into articulatory errors, based on place of articulation, manner, voicing, and vowel quality articulatory deviations. The experimental analysis performed on the non-native English speech had a phoneme recognition accuracy of 91.4 that was much higher than the commercial ASR-based systems (78.3) and the traditional HMM-GMM baselines (70.5). The system was very sensitive to ESL pronunciation errors, making it 84 percent accurate in substitution, 82 percent accurate in deletion and 79 percent accurate in insertions in detection and articulatory mapping was over 87 percent accurate in all categories. The framework was tested in Python with deep learning packages and speech processing toolkits, and provided a scalable, explainable, and learner-focused system that can be used to support the intelligent training of ESL pronunciation and provide pedagogically significant feedback at the phoneme level.**

Keywords*—Articulatory error analysis; attention mechanism; ESL pronunciation assessment; phoneme recognition model; speech processing framework

I. Introduction

English has emerged as a lingua franca all over the world and therefore, oral fluency is paramount in academic and professional spheres [1]. Nevertheless, ESL students still perform poorly especially at phoneme level, where small articulatory errors can result in a large amount of intelligibility loss. The students who speak ESL can hardly pronounce, intonate and stress correctly at the phoneme level, as the feedback provided on the pronunciation in the classroom is not always effective in assisting with correction and confidence development at the individual level [2]. As a result, technology-based learning tools are becoming the popular solutions to ESL pedagogy. The most sought out systems are those that can deliver individualized real-time, and objective feedback. Such kinds of tools do not only increase the support to more groups of learners, but it also helps in developing the ability of learners to support spoken English in a more irregular and measurable way. In the case of the ESL learners, it is important to note that their pronunciation is a major factor influencing their communication abilities that may help them interact with the native and non-native speakers effectively [3]. Most ESL learners still struggle with phoneme-level issues, such as pronunciation, intonation, and stress, despite advances in the fields of grammar and vocabulary acquisition [4]. In conventional classroom practice, it is difficult to give each student individual pronunciation feedback because of time limitations and the subjectivity of teacher feedback [5]. Therefore, the interest in the incorporation of technology-based solutions into the sphere of ESL pedagogy is increasing, especially in the case of the solution that will provide real-time, objective, and individual feedback that can contribute to the enhancement of the spoken English proficiency of the learners.

Over the past few years, there have been a great many speech recognition and pronunciation evaluation tools available in the language learning and practice environment [6]. The general result is generic or very broad feedback that will not help the student to make accurate articulatory adjustments. The absence of interpretability in the feedback also keeps learners from taking action based on the results. Fluency, accuracy and rhythm are other parameters that some sophisticated tools use scoring algorithms to rate [7]. Nevertheless, the majority of existing models focus more on accuracy at the word or sentence level as compared to accuracy at a phoneme level [8]. Besides, most of these systems have been trained using native speech more often than not and thus prove to be less effective during the analysis of non-native speech patterns, a characteristic of ESL learners [11],[9]. Moreover, little emphasis has been made to explain why feedback should be, and explainability has made the learners insecure about rectifying their errors. Such gaps require a more advanced AI-powered solution that can provide more accurate, actionable, and learner-friendly information about pronunciation [10]. To overcome these drawbacks, a new Articulatory Error Mapping Engine, which transforms phoneme-level mispronunciations into place-, manner-, voicing-, and vowel-quality-feedback, is proposed in this study to provide interpretable, accurate and pedagogically relevant pronunciation correction to ESL students.

A. Problem Statement

Moreover, a significant number of pronunciation evaluation tools use native-speaker training data, which makes them less effective when they are implemented to accented ESL speech. This drawback covers the precise identification of the typical phoneme-level mistakes like substitutions, deletions, and insertions, especially of the sounds that are acoustically similar. Even those systems which present phoneme-level feedback are often linguistically uninterpretable, and present numerical scores or visual feedback which cannot be accessed by a novice learner without advance phonetic knowledge [12]. Moreover, there exist no systematic articulatory explanation, which restricts the pedagogical usefulness of the available solutions, since the learners very often do not know how articulation features like place, manner, voicing, or vowel quality can lead them to their mistakes[13]. As a result, there is an evident necessity of a framework of pronouncing measurements not only reaching to the reliability of phoneme-level recognition of accented ESL speech but also converting the identified mistakes to the articulatory feedback that can be deciphered and helps guide teaching and a quantifiable progress [14].

B. Research Motivation

The rationale behind this research is due to a conceptual weakness of current ASR-based pronunciation assessment systems; the disenfranchisement between acoustic recognition performance and semantically communicative feedback. Although it can be seen that most deep learning architectures can enhance the accuracy of phoneme classification, not much attention has been given to the correspondence of internal phoneme representations with articulatory behavior. The majority of existing systems consider the prediction of phonemes as an acoustic optimization problem, and they do not incorporate structured articulatory knowledge into the model process. The phoneme representations taking the form of articulatory-aware supervision is the concept that allows the presentation of not only spectral and temporal patterns but also organised linguistic properties like place of articulation, manner, voicing and vowel quality. This view changes the score-based assessment of pronunciation to the interpretation-based representation. The framework attempts to address a root-deficiencies in ESL pronunciation learning systems by connecting both acoustic modeling and articulatory reasoning as a means to deliver to student's feedback that would be technical and pedagogical in their actions.

C. Research Significance

This research will solve one of the inherent shortcomings of current ESL pronunciation measurement systems, which is the lack of connection between the accuracy of phoneme recognition and the feedback that is useful to pedagogical purpose. Whereas a lot of models have been based on optimizing the performance of classification systems, little has been done in determining the manner in which the representations of phonemes can be organized in terms of articulatory behavior. The suggested structure presents articulatory-conscious modeling, in which the phoneme deviations are identified and explained in a systematic way by structured articulatory classes. The framework fills the gap between acoustic modeling and linguistic explanation by introducing articulatory properties to the feedback mechanism, allowing to accurately detect but also interpret errors that have a cognitive meaning. This change in the process of score-based assessment to more structured articulatory reasoning improves self-correction in the learner, facilitates instructor diagnosis, and improves the interpretability of AI-assisted pronunciation assessment systems.

D. Key Contributions

- The proposed study presents a phoneme-based paradigm used in studying ESL pronunciation mistakes where fine articulatory deviations are considered in the paradigm, which is normally ignored when the word or sentence level is used in analyzing them.
- It presents a systematic approach to phoneme alignment and representation, which makes it possible to analyze the non-native English speech in detail at the level of sub-words and study the pronunciation of the word under discussion.
- The study establishes a scheme of systematic articulatory error representation which aligns the presence or absence of phoneme deviations with the categories of place of articulation, manner of articulation, voicing and vowel quality.
- A feedback mechanism, which can be linguistically interpreted, is created to convert phoneme-level errors into articulatory advice that is easily comprehensible and easy to act upon, to make pronunciation assessment systems more pedagogically useful.
- The ESL-centered pronunciation assessment model is authenticated to effectively detect phoneme replacements, additions, and removals in order to be used to facilitate specific teaching and quantitative pronunciation growth.

E. Rest of the Section

Section II conducts literature research on connected work and AI-based methods of pronunciation evaluation in ESL. Section III outlines the proposed hybrid methodology, which includes data processing, alignment of phonemes, feature extraction and CNN-BiGRU-Attention architecture. Section IV presents the experimental findings, measures of evaluation, analysis of errors and comparative researches. Section V comes up with the main findings, focusing on the detection of phoneme level errors, customized feedback and research directions.

II. Literature Review

Abimanto and Sumarsono [15] discusses how the application of AI can be integrated into the Google Read Along application to aid struggling learners in improving their English pronunciation. This study will use a quasi-experimental research design, and there will be 70 students in this research (one group control and the other experimental). The N-Gain test is used to determine pronunciation gains, of which the current N-Gain test results indicate an improvement, when used with Read Aloud-AI. The positive effect of real-time AI-based error correction has been confirmed with the feedback obtained in questionnaires and interviews. The findings indicate that the AI-based Read Aloud produced an effective and viable pronunciation intervention tool among learners of ESL.

Xu [16] investigate the usefulness of deep learning to assess the quality of English pronunciation, especially on ASR tool, where the learners experience limited exposure to the language. Major aspects considered in the study are intonation, speed, and rhythm, where machine analysis is compared to hand-marking. An analysis of 240 instances reveals that the deep learning model has significant reliability, and 100 percent accuracy does not differ much from the human evaluation. The system also facilitates the learner in determining the gaps in the pronunciation by comparing the pronunciation with the standard one frequently. The conclusions raise the opportunities of deep learning to increase accuracy, objectivity, and efficiency levels of English pronunciation assessment.

Hoang, Han, and Le [17] assesses how Mission Fluent AI chatbot can help vocational students in Hanoi improve their pronunciation of English. The study is based on a quasi-experimental design to learn that the experimental group had better pronunciation than the control group, based on the use of 60 participants. The results of the survey and interview feedback indicate the remarkable engagement level of learners and positive attitude towards chatbot-assisted instruction. However, even the best of them has practical difficulties that the study has recognized as the cost and implementation limitations. The conclusions underline the opportunities of AI chatbots to contribute to improving pronunciation skills in vocational education.

Raja and Sanghani [18] explore the field of SER that attempts to extract features like pitch, tone, and intensity to label emotions in speech. The article points out the application of different classifiers in identifying emotional difficulties such as happiness, sadness, anger, and neutrality. Although there is an increasing presence of SER datasets, the problem is still stark because of the subjectivity in emotions and the inability to annotate them well. The researchers highlight the increasing applicability of SER in the speech processing context. Their contribution supports the requirement for finer tools and algorithms to enhance the robustness of the speech classes using emotions.

Permatasari [19] evaluate how the ELSA Speak app can be effective in enhancing the pronunciation abilities of EFL students in a quasi-experimental research. When comparing the results using pre-test and post-test design in the sample of ELSA Speak and U-Dictionary users, the groups of both the experimental and the control group differ significantly and it is statistically significant. The study employed paired sample t-tests in the determination of the effectiveness of each tool. The findings suggest that the ELSA Speak is more successful in pronunciation acquisition regarding the EFL usage. This elicits its possible usefulness as a pedagogical aid in language learning programs in enhancing speaking skills.

Babaeian [20] points out the importance of pronunciation in effective communication and ineffectiveness of human rated judgments, such as inconstancy and subjectivity. These systems assess both the suprasegmental and segmental features in order to provide ESL learners with more accurate feedback. Despite apparent benefits, issues of model validity, data diversity, and ethics are mentioned in the article as well.

Zaveri [21] introduced the Transformer-based with large Audio Models like HMM-GMM Baseline to access the automatic Speech Recognition, Text-to-Speech, and Music Generation. Models such as SeamlessM4T can now support multilingual and multi-task processing of speech without having to use task-specialized systems. This paper introduces a long-term summary of the state-of-the-art techniques, on-the-job applications, performance indicators, and constraints existing. It also defines the direction of future research as well as offers a dynamic source of open-source implementations so as future developments in audio AI can be made.

Zou et al. [22] explore the effectiveness of the AI-based speech evaluation supplemented with pronunciation visualization tools in the purpose of the Chinese EFL students studying at the university level. The intervention was done with a group of 50 students based on a semi-automated system that provides both, waveform and spectrogram feedback in 8 weeks. Along with the improvement of learners self-monitoring and significant segmental accuracy, significant segmental accuracy improvement turned out to be the results. Learners reported about the improved awareness of articulation patterns and error correction. However, the study indicates that visual data decoding requires previous training in phonetics and, therefore, can be limited to beginners.

Kheir, Ali, and Chowdhury [23] investigate the use of deep convolutional neural networks to assess segmental and suprasegmental characteristics in Arabic-speaking ESL learners. Data to train their system was on annotated speech, and their feedback was on a phoneme-level with real-time scores. The findings indicated that there were higher rates of engagement among the learners and a higher rate of accuracy in pronunciation by 16 percent in 6 weeks. Real-time feedback and high scalability are strengths, but rare phoneme sequences and accentual variation were a challenge to the system.

Yoo and Ahn [24] examine how AI-based prinunciation tutors can provide diverse feedback to Korean EFL students. The study evaluates the performance of an intelligent tutor system that is able to automatically recognize phoneme errors and generate correction plans based on articulation specifics. The system was shown to have a great improvement in consonant cluster production, based on a sample of 60 learners. Despite the fact that students liked the concrete feedback, technical problems that came with spontaneous speech recognition were noted in changing the pace of speech and noises. The literature review summary is illustrated in Table I.

TABLE I. DATASET SUMMARY TABLE

Attribute	Description
Dataset Name	Mozilla Common Voice
Data Type	Read speech with textual transcriptions
Validation Method	Crowd-sourced listener voting
Subsets	Train, Development, Test
Audio Format	MP3
Metadata	Age group, gender, accent, votes
Speaker Accents	Multiple native and non-native English accents
Accessibility	Publicly available

The literature review supports the fact that AI-based pronunciation tools have obvious advantages to language learning, yet it indicates significant shortcomings as well. The majority of current systems are based on word- or sentence-based scoring and cannot detect the error at the phoneme level and provide the reasons why it takes place. Most of them are trained on data of native speakers and are therefore ineffective with accented ESL speech and some give feedback with no articulatory insight. The current paper fills these gaps and opens a Phoneme-level recognition to place, manner, voicing, and vowel-quality features with the introduction of an Articulatory Error Mapping Engine. By providing a clear, linguistically-based, and learner-focused corrective feedback, this method allows the close correspondence to the actual ESL pronunciation requirements.

III. PROPOSED FRAMEWORK ON AI-BASED PHONEME RECOGNITION FOR ESL ENGLISH PRONUNCIATION EVALUATION

The given framework presents an articulatory-conscious phoneme modeling approach, which combines the learning of the acoustic features and organized linguistic comprehension. In contrast to the traditional pronunciation assessment pipelines, which consider phoneme recognition and feedback productions as two separate processes, the given methodology creates a direct linkage between the phoneme prediction and the articulatory reasoning. Multi-stage pipeline is constructed in such a way to identify phoneme level deviations in ESL speech as well as to organize the deviations on the basis of articulatory dimensions that can be interpreted.

Its architecture is a series of modules such as data preprocessing, phoneme alignment, acoustic-prosodic feature extraction, hybrid CNN-BiGRU-Attention-based phoneme modeling, and Articulatory Error Mapping Engine. The system uses spectral convolutional modeling, bidirectional temporal learning and attention based contextual weighting to identify finer-grained patterns of pronunciation and remain interpretable. The articulatory mapping step converts the outputs of the classification step into place-, manner-, voicing-, and quality of the vowel-based explanations, therefore permitting linguistically based corrective feedback. The proposed framework is shown in Fig. 1.

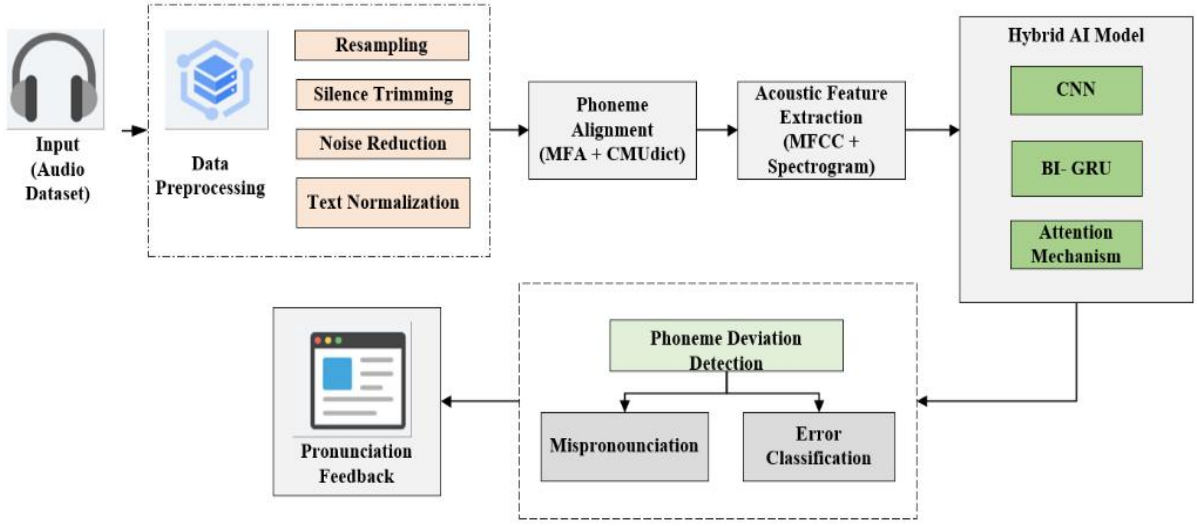

Fig. 1. Proposed framework.

A. Data Collection

The speech text in this research was taken in the Mozilla Common Voice corpus which is large-scale publicly available speech corpus devoted to the investigation and fine-tuning of automatic speech recognition systems. Audio samples are authenticated by means of a voting system with recordings that are considered valid being given majority approval by more than one listener. The dataset is divided into development, training, and testing subsets and allows systematic models to be developed and evaluated. The metadata of each audio clip such as the age group, gender, and the accent of the speaker are

provided, which makes it easier to select non-native English speech. Audio files are stored in a compressed format and have these metadata files, which make it useful to perform phoneme-level alignment and a pronunciation analysis [25].

Table provides the description of the Mozilla Common Voice dataset, which is a dataset of speech, its type, the method of its validation, partitions of the dataset, audio codec, speaker information, the variety of accent and the appropriateness of this dataset in the study of phonemic pronunciation analysis in ESL-oriented speech processing research.

B. Data Preprocessing

The audio and transcribed information is pre-prepared to be able to obtain quality input which can be analyzed to the phoneme level. The audio files are then resampled to common 16 kHz sampling rate and this makes the dataset at the same frequency. It is then followed by silence trimming which is done with the help of an energy threshold to eliminate non-verbal pauses that may disrupt the proper phoneme segmentation. By spectral subtraction, the background noise is removed and the speech signals are made clearer to make them easier to analyze. A normalization of the text on the transcript side is done to convert all text to lowercase, eliminate punctuation and clean up tokens to fit with the phoneme dictionary. This preprocessing provides a very exact, one to one correspondence between what is said and what is phonemically represented. The resulting clean and aligned inputs give a stable and consistent starting point of the phoneme alignment, feature extraction and model inference. Fig. 2 also shows the particular steps and the flow of data preprocessing to complete each step of raw input to the complete data to be used in the pronunciation assessment structure.

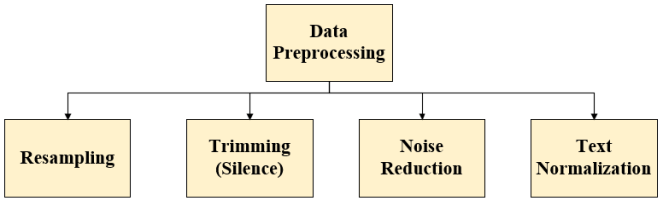

Fig. 2. Data preprocessing steps.

1) Resampling: Resampling involves the process of changing all the audio signals to a compatible sampling rate. To facilitate calculations of the feature extraction and alignment modules, all recordings are resampled to 16 kHz and the equation for it is represented in Eq. (1),

$$y[m] = x\left[\frac{f_s}{f_t}.m\right] \quad (1)$$

where, $y[m]$ is resampled signal, f_s is original sampling rate, f_t target sampling rate, m resampled time index.

2) Silence trimming: Silence trimming is to cut low-energy parts on both the head and tail of the audio to conduct analysis on speech. It is especially necessary to ESL learners who tend to place long pauses in speech. The energy of the frame is computed in Eq. (2),

$$E_i = \sum_{n=1}^{N} |x_i[n]|^2 \quad (2)$$

where, E_i is the energy of the frame, $x_i[n]$ is audio sample at time, N is number of samples in a frame.

3) Noise reduction: The reduction of noise improves the speech signal by eliminating the background sounds. Spectral subtraction has been used in this study to ensure the impact of the ambient noise is reduced, which is even prevalent in crowd-sourced recordings such as Common Voice. The speech spectrum is represented in Eq. (3),

$$\hat{S}(f) = |X(f)| - |N(f)| \quad (3)$$

where, $X(f)$ is observed noisy speech spectrum, $N(f)$ estimated noise spectrum, $\hat{S}(f)$ is enhanced speech spectrum.

4) Text normalization: Text normalization reads up the transcripts by converting them into canonical form. Non-verbal tokens, punctuation and casing are eliminated. The normalized sequence is ontained from Eq. (4),

$$T' = \{lowercase(t_i) | t_i \in T\} \quad (4)$$

where, T is original transcript token list, T' normalized token list, t_i is individual token in transcript.

C. Phoneme Alignment and Mapping

The phoneme alignment is necessary to define a time correspondence between the learner speech and phonetic transcription of the speech in question so as to obtain a clear segmentation of the constant speech into the units of phonemes. This makes it easy to undertake phoneme level analysis and detection of errors. The alignment process in this research has three major steps including pronunciation dictionary look up, forced alignment with MFA, and phoneme boundary extraction and formatting.

1) Pronunciation dictionary lookup: A traditional pronunciation dictionary is employed to convert the normalized transcript to a sequence of phonemes. The CMUdict which provides the phonetic pronunciations of the English words in ARPAbet format. Each transcript word's pronunciation must be looked up, and the phonemic correspondence will be used. The phoneme sequences will act as a benchmark to which the learner will be guided regarding his or her speech. Given that the phone sequence is based on dictionaries, the system is reliable and linguistically valid. It is expressed in Eq. (5).

$$P = \{p_1, p_2, p_3, \ldots, p_n\}, \quad \text{where } p_i = \text{CMUdict}(w_i) \quad (5)$$

where, w_i in the transcript is mapped to its corresponding phoneme p_i using a pronunciation dictionary. The resulting sequence P serves as a canonical reference to compare with the pronunciation of the learner at the end making everything consistent and linguistically correct in the phoneme representation.

2) Forced alignment using montreal forced aligner: The phonemes are then paralleled with the speech of a learner in the form of the Montreal Forced Aligner (MFA) after the creation of the phoneme sequence (factor of dialect was introduced later). MFA provides a more recent method of aligning audio with phonetic transcripts, making guesses of the context of

every audio fragment and setting accurate start and stop times of every phoneme. This timed division is especially applicable in the ESL learners as the system is able to detect weaknesses in pronunciations at the phoneme level. Both acoustic and temporal structure of speech are captured by the alignment and the deviations in the pronunciation are correctly analyzed. The process of alignment is shown in Eq. (6):

$$A = \{(p_i, t_i^{start}, t_i^{end})\}_{i=1}^{n}, \quad t_i^{start}, t_i^{end} = \text{MFA}(p_i, x(t)) \quad (6)$$

Where alignment process Amaps each phoneme $p_i$to its start (t_i^{start}) and end (t_i^{end}) time in the learner's audio signal $x(t)$.

3) Phoneme boundary extraction: The alignment process produces phoneme boundaries containing the start and the start time of a TextGrid. The parts are examined one by one with the aim of recovering acoustical attributes and examining the anticipated phonemes to the canonical sequences, and discovering the effectiveness of procured articulation. This enables a fine grained capture of wrongly pronounced phonemes, including insertions, omissions or replacements and also gives one on one feedback on pronunciation with visual emphasis on wrongly pronounced phonemes. It is expressed in Eq. (7)

$$B = \{(p_i, t_i^{start}, t_i^{end}) \mid i = 1,2,\dots,n\}, \quad \text{stored as TextGrid} \quad (7)$$

where the phoneme boundaries Bprovide start and end times for each phoneme p_i, allowing precise extraction of phoneme-specific acoustic features. This structure facilitates detection of pronunciation errors like insertions, deletions, or substitutions and enables personalized feedback for ESL learners.

D. Acoustic Feature Extraction

Once the phoneme alignment is done, the second step is the conversion of each segment of the text grid into a structured representation of the acoustic properties of the speech referred to as the feature extraction stage. This is a very important step because it gives the neural network a chance to get the data to differentiate between the right and the wrong phoneme pronunciations. Each of the time-linked phoneme records a mixture of low- and mid-level characteristics such as log-Mel spectrograms, MFCCs, and prosodic characteristics such as pitch, duration, and energy. Such extracted features are subsequently fed into the hybrid deep learning system, which allows proper phoneme categorization and assists with identifying the pronunciation mistakes, as explained in Eq. (8):

$$F_i = \{LogMel(x_i), MFCC(x_i), Pitch(x_i), Energy(x_i), Duration(x_i)\} \quad (8)$$

where each phoneme segment $x_i$is converted into acoustic features F_i , combining spectral, prosodic, and temporal properties.

1) Log-mel spectrograms: The log-Mel spectrograms are the energy of the speech over time and frequency but using the perceptually-based Mel scale which puts more emphasis on the low frequencies and squeezes the high frequencies. They are produced as the result of short-time Fourier transform, Mel filtering, and log compression and represent formants and articulatory transition. CNN layers are trained on spectral patterns, phonetically similar sounds, to be distinguished to conduct an appropriate analysis of ESL pronunciations. It is calculated in Eq. (9):

$$S_{log\_mel}(t,f) = \log\backslash Big(\sum_{k=1}^{K} |X(t,f_k)|^2 \cdot H_{mel}(f_k,f)\backslash Big) \quad (9)$$

where the Log-Mel spectrogram is the distribution of energy of speech on a perceptual Mel scale. It records phoneme changes, formants and articulation patterns. Convolutional layers are able to learn local spectral patterns, which can distinguish between phonetically similar-sounding in pronunciation of ESL learners.

2) Mel-frequency cepstral coefficients: MFCCs isolate the spectral perfection of speech, which is an indication of voice tract resonances. They are frequency content compression methods extracted to cover both the frequency content and are resistant to pitch and volume changes through Fourier transform, Mel filtering and DCT. First and Second derivatives are used to calculate temporal dynamics. The model can be fed with the MFCCs into recurring layers so as to analyze changes in articulation over time. They are particularly successful at identifying the slight phonetic variation and mispronunciation, which would offer a strong framework on which to judge the phoneme production of ESL students and could be used in providing accurate feedback on pronunciation. It is mentioned in Eq. (10):

$$\text{MFCC}_n = \sum_{m=1}^{M} \log\backslash Big(\sum_{k=1}^{K} |X(t,f_k)|^2 H_{mel}(f_k,f_m)\backslash Big)\cos\backslash Big\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\backslash Big\right] \quad (10)$$

where MFCCs code spectral envelope of speech on the basis of Fourier transform, Mel filtering, and DCT. Temporal derivatives are volatile. They give good abilities of detecting the faintest phonetic changes and help in proper performance of pronunciation in ESL students.

3) Temporal and prosodic parameters: Rhythm, stress, and fluency are represented in the form of the temporal and prosodic features. Phonemic length determines the existence of overlong vowels, or absent consonants, the intonation of the intonation and stress placement is detected through pitch, and the force of articulation is detected through energy. Taken together with spectral properties, these features make a detailed picture on a phoneme level. Log-Mel spectrograms, MFCCs and prosodic measures are normalized and presented to the hybrid neural network in form of tensors. It can be used to check pronunciation errors (like mistimed articulation, improper tongue or lips position, and stress misplacement) which are the basis of phoneme classification and individualized feedback. It is equated in Eq. (11):

$$F_{prosody} = \{\text{Duration}(p_i), \text{Pitch}(p_i), \text{Energy}(p_i)\} \quad (11)$$

In the cases of phonemes duration, pitch and energy as prosodic features. Furthermore, these features in conjunction

with spectral features determine rhythm, the level of stress and fluency. They are effective in the detection of mistimed articulation or wrong stress on a word or weak articulation. Finally, the features form the foundation of phoneme classification and offer individual pronunciation feedback to every person.

Algorithm 1 explains the acoustic Feature Extraction provides the steps involved in transforming raw audio signal into feature vectors to analyse speech. The system removes Mel-Frequency Cepstral Coefficients (MFCC) as spectral representation, log-Mel spectrograms as frequency-time patterns and temporal representations, as phoneme duration, pitch, and energy, depending on the type of feature selected. The various types of features can be used in combinations to enhance richer representations such as MFCC with spectrogram, or MFCC, spectrogram, and temporal features. Extracted features are combined into a unified vector that is used as the neural network model input in such tasks as phoneme recognition and pronunciation error detection.

Algorithm 1: Acoustic Feature Extraction

```
Input: Raw audio signal
Output: Feature vector (MFCC / Spectrogram / Temporal / Combination)
  if feature_type == 'MFCC':
        Extract Mel-Frequency Cepstral Coefficients
        features = extract_mfcc(audio)
  if feature_type == 'Spectrogram':
        Extract log-Mel spectrogram
        features = extract_spectrogram(audio)
  if feature_type == 'Temporal':
        Extract temporal features like duration, pitch, energy
       features = extract_temporal_features(audio)
   if feature_type == 'MFCC+Spectrogram':
            Combine MFCC and Spectrogram features
        mfcc = extract_mfcc(audio)
        spec = extract_spectrogram(audio)
        features = concatenate (mfcc, spec)
  if feature_type == 'MFCC+Spec+Temp':
       Combine all features
            mfcc = extract_mfcc(audio)
            spec = extract_spectrogram(audio)
            temp = extract_temporal_features(audio)
            features = concatenate (mfcc, spec, temp)
   else
            Invalid feature type specified.
            features = None
 return features
```

E. Phoneme Recognition on Hybrid Deep Learning Module

The main premise of the suggested structure is a deep learning hybrid architecture that will be used to determine phoneme-level pronunciation patterns of ESL learners. The model takes acoustic features such as log-Mel spectrograms, MFCCs and prosodic parameters as inputs (and predicts labels of phonemes). Its structure consists of four parts, including CNN, Bi-GRU, attention mechanism, and output classification layer, which are used to detect fine-grained errors based on time-aligned segments of phonemes. The results of these are then inputted into the Articulatory Error Mapping Engine, which gives the results in a linguistically interpretable and pedagogically valuable feedback. Fig. 5 demonstrates the Hybrid CNN-BiGRU-Attention model of phonemes recognition.

1) Convolutional Neural Network: CNN layers process the 2D Log-Mel spectrograms and as a result respond to spatially localised events, such as formants, frequency modulations and energy shifts.

2) Bidirectional Gated Recurrent Units: The Bi-GRU layers then extend the CNN layers to capture temporal relationships and discover sequence patterns as they read the input sequences forward and backward and get individually used with ESL learners and offer the opportunity to recognize phonemes in continuous speech.

3) Attention mechanism: Bi-GRU generates an attention layer, which is subsequently implemented to weight the importance of every time or phoneme. It places additional emphasis on mispronounced phonemes when implemented among ESL learners making the model more discriminative and enabling more intelligible and focused detection of pronunciation deviations in a sentence.

4) Output classification layer: This is the last layer and is a fully connected softmax that provides the probability of phonemes (on a time-aligned segment) probability. Segments are given predicted labels due to the most likely occurrence and compared to canonical sets of phonemes. The output also drives the pronunciation feedback module allowing the pronunciation accuracy to be evaluated and a specific corrective feedback to be generated to work with ESL learners.

F. Phoneme-Level Error Processing

The last stage of the system produces learner specific feedback through comparison of predicted phoneme sequences and reference sequences. It finds pronunciation errors on a phoneme level, such as substitutions, omissions and insertions using alignment metadata and classification output. The individual errors associated with each linguistic category (voicing, place of articulation, and so forth) are mapped to a feedback template. The system generates clear textual feedback, which shows the particular phoneme and provides articulatory corrective advice. The errors are then run through the Articulatory Error Mapping Engine that translates them into articulatory categories that can be interpreted to provide learners with accurate, actionable and pedagogically significant feedback. Fig. 3 shows hybrid AI-framework.

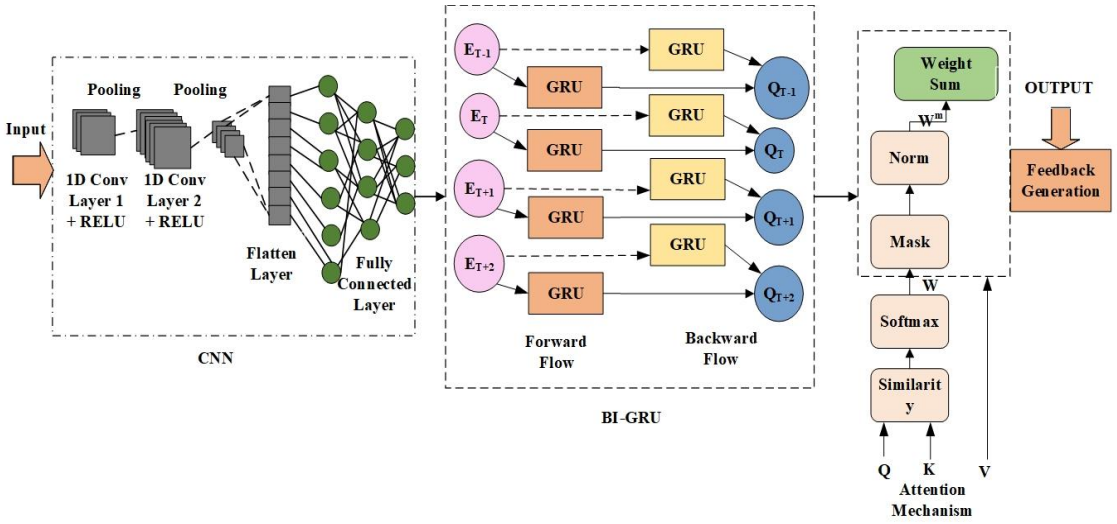

Fig. 3. Hybrid AI framework.

G. Articulatory Error Mapping Engine

The Articulatory Error Mapping Engine is the fundamental novice of the suggested framework. Although the hybrid CNN-BiGRU-Attention model is capable of recognizing phoneme-level substitutions, insertion, and deletion errors, they need to be linguistically interpreted so that they can be pedagogically significant to ESL learners. The phoneme that is mispronounced is mapped to its articulatory characteristics such as place of articulation, manner of articulation, voicing, and vowel quality to a predefined phoneme-to-articulatory dictionary. The errors are identified and registered as errors in articulation (e.g., the alveolar/dental, voiced/voiceless, high/mid vowel, etc.). The engine finds the corrective feedback templates associated with the non-canonical articulatory feature so that actionable advice is provided, e.g. production shifted to alveolar /t/; /th/ needs the tongue between the teeth. This mapping converts raw classification results into linguistically intelligible feedback (with a learner-oriented result), which is the difference between the system and other conventional ASR-based pronunciation evaluation tools. The flow-chart of the phoneme recognition and feedback generation process run by AI is presented in Algorithm 2.

Algorithm 2: AI-Based Phoneme Recognition and Feedback Framework

BEGIN
1. Data Collection
    Load ESL speech samples from Mozilla Common Voice
    Select samples tagged with non-native accents
2. Data Preprocessing
    FOR each audio sample
        Resample audio to 16 kHz
        Apply silence trimming using energy threshold
        Perform spectral subtraction for noise removal
        Normalize transcripts (lowercase, remove punctuation)
3. Phoneme Alignment
    FOR each transcript
        Convert text to phonemes using CMU Pronouncing Dictionary
        Align phonemes using Montreal Forced Aligner (MFA)
        Save phoneme boundaries with timestamps
4. Feature Extraction
    FOR each aligned phoneme segment
        Extract Log-Mel spectrogram
        Extract MFCC + delta + delta-delta
        Extract prosodic features (duration, pitch, energy)
5. Phoneme Classification
    FOR each feature vector
        Apply CNN for spatial feature extraction
        Pass output to Bi-GRU for temporal modeling
        Apply attention to emphasize key phoneme frames
        Predict phoneme label using softmax layer
6. Error Identification
    FOR each predicted phoneme
        IF predicted phoneme == reference phoneme
            Mark phoneme as correctly pronounced
        ELSE
            Classify error as substitution, deletion, or insertion
7. Articulatory Error Mapping Engine
    FOR each mispronounced phoneme
        Retrieve articulatory features of reference phoneme
        Retrieve articulatory features of predicted phoneme
        Identify deviation in:
            - Place of articulation
            - Manner of articulation
            - Voicing
            - Vowel quality
        Select appropriate corrective feedback template
        Generate articulatory explanation (e.g.,
        "Tongue should be between teeth for /θ/, but moved to alveolar /t/.")
8. Output Generation
    Display or export structured feedback combining:
        - Phoneme-level correctness
        - Error type
        - Articulatory deviation
        - Corrective instruction
END

Algorithm 2 describes the AI-Based Phoneme Recognition and Feedback Framework of systematic pronunciation assessment. Preprocessing of non-native speech samples is carried out by resampling, silence cutting, noise cancellation and transcript normalization. MFA is the conversion of words to phonemes and alignment. Each segment is extracted and features (spectral, MFCC and prosodic) are computed (duration, pitch, energy). Errors in pronunciation are detected by the Articulatory Error Mapping Engine, which transforms it into a place, manner, voicing, or vowel-quality error. The system then develops learner-centered, structured feedback on pre-defined articulatory templates to give interpretable and pedagogical useful pronunciation instructions.

IV. Results and Discussion

The suggested framework became very successful in identifying patterns at phoneme level in the speech of ESL speakers and properly identifying some common mistakes in pronunciation, such as substitutions, deletions, or additions. This was able to differentiate closely related phonemes and record articulatory deviations which are normally difficult among non-native speakers. The system empowered articulatory categories by incorporating the Articulatory Error Mapping Engine to give the recognition outputs clear articulation categories, and it provided articulate explanations of errors. This interpretability made the pedagogical value of it more precise and learner-centered. In general, the findings show that the framework provides correct phoneme recognition, high-quality error detection, and linguistically based corrective support thus is an excellent tool to be used in practical ESL pronunciation training.

A. Experimental Setup

The experiment took the Common Voice dataset, which was phonemically synchronized by MFA and CMUdict. processed audio and extracted features, such as MFCCs, spectrograms, and temporal features. The Python-based model was tested in terms of accuracy, precision, recall, and F1-score. The parameters of the simulation are shown in Table II.

TABLE II. Simulation Parameters

Parameter	Value / Description
Sampling Rate	16 kHz
Window Size	25 ms
Window Stride	10 ms
Number of MFCCs	13
Spectrogram Type	Log-Mel Spectrogram
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	50
Dropout Rate	0.3 (applied to Bi-GRU and FC layers)
Activation Function	ReLU (CNN), Tanh (Bi-GRU), Softmax (output)
Loss Function	Categorical Cross-Entropy

B. Phoneme Recognition Performance

Phoneme recognition module was introduced on the Common Voice English corpus, a hybrid CNN-BiGRU-Attention architecture was trained on a strictly selected subset of the corpus, which was specifically chosen to reflect the non-native English pronunciation. The phoneme sequences were phoneme sequences that were synchronized through forced alignment to enable accurate timing matching of predicted and reference phonemes. Table II illustrates the performance indicators on phoneme recognition.

TABLE III. Phoneme Recognition Performance Metrics

Metric	Value (%)
Accuracy	91.4
Precision	89.7
Recall	90.1
F1-Score	89.9

The final model that combines AI and a hybrid phoneme recognition model performed well enough over all of the evaluation measures, giving a balanced performance on all the metrics. The model was highly accurate in terms of identifying phonemes uttered by ESL learners, given that the overall accuracy was 91.4%. These findings point to the fact that the system has a high level of reliability and can be used in ESL to test the articulation of phonemes of pronunciation. Fig. 4 shows phoneme-level accuracy plot.

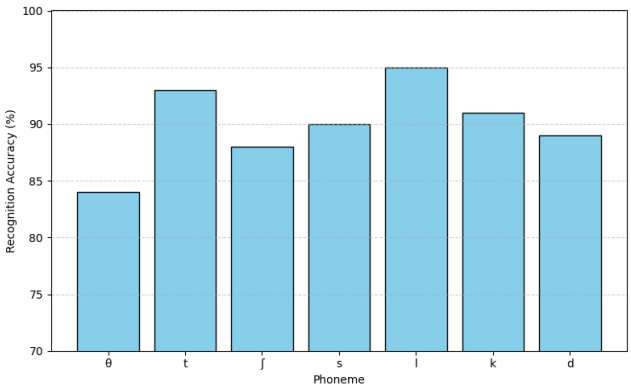

Fig. 4. Phoneme-level accuracy plot.

The proposed hybrid AI model recognizes individual phonemes by 56.4 % accuracy as shown by Fig. 6. The highest accuracy score was achieved by /l/ phoneme at 95 %, then there was a close-up of /t/ and /k/ which means that the consonantal sounds that are dominant were well performed. More difficult phonemes, which ESL learners are usually unable to pronounce well, still achieved 83.95 and 88.33 score. In general, the figure brings to the fore the fact that the model in question demonstrates a consistent phoneme-level classification of various ESL speech. The predictions are entered into the Articulatory Error Mapping Engine, where the wrong pronunciation can be translated in terms of its particular articulatory error.

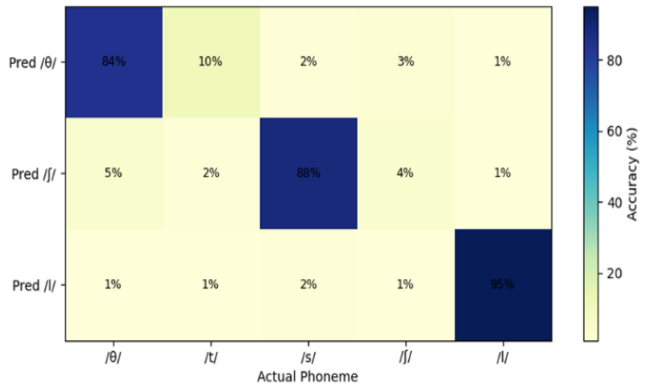

Fig. 5. Phoneme confusion matrix.

Fig. 7 explains the accuracy of the model to sort out the similar phonemes in ESL speech. The rates of high accuracy are noted by /theta/, /sh/, and /l/, which show models' recognition at the scale of 84%, 88%, and 95% respectively, suggesting an excellent precision level of models. Another thing that is noticed in the matrix is that /theta/ is sometimes wrongly categorized as /t/ (10%) and /sh/ as /s/ (4%), something which is in line with common ways in which ESL students are mispronouncing. In predicting /l/, minimal confusions are evident, demonstrating the effectiveness of the system on considerations that are late.

C. Error Detection and Categorization

Deletion errors are errors that are committed when a phoneme is not pronounced. Through the comparison of the aligned predicted phoneme sequence with their corresponding reference phoneme sequence, the system is capable of detecting such errors. The accuracy of detection of each type of error was calculated individually in order to determine sensitivity to the various articulation patterns. Table III shows a summary of the detection accuracy of all error categories.

TABLE IV. DETECTION ACCURACY

Error Type	Detection Accuracy (%)
Substitution	84
Deletion	82
Insertion	79

The system proved to be quite successful in identifying all three large types of pronunciation errors among the ESL learners. The substitution errors were found to be detected with the best accuracy rate of 84%, which is a sign of the sensitivity of the model towards the frequent mispronunciation of phonemes. The deletion errors were also well detected with an 82 percent detection, so it could be said that error deletion could be reliably detected. Insertion errors were also found in slightly lesser precision of 79% however, they were still spotted quite accurately.

Fig. 8 shows the number of errors successfully detected by three significant error types, namely substitution, deletion, and insertion. The highest accuracy rate of 84% was achieved in identifying errors involving replacements of phonemes which are characteristic of ESL speech, revealing the strength of the model. In the next position, deletion errors existed at 82%, which indicates consistent results in sound omission detection. Insertion error, which was a little bit lower at 79%, was adequately identified, thus showing overall model competence. The high credibility of the phoneme-level error analysis by the system is highlighted by similar error detection rates in all the error types.

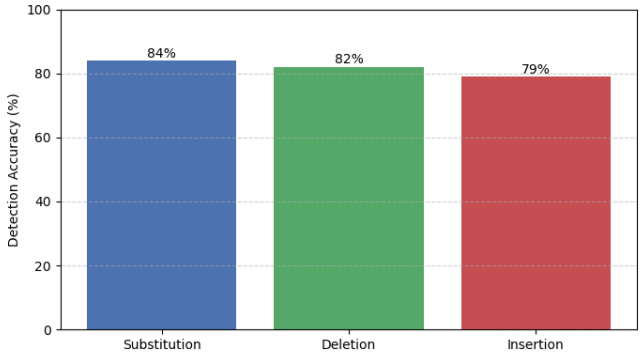

Fig. 6. Detection rate by pronunciation error.

D. Effectiveness of the Articulatory Error Mapping Engine

To determine the pedagogical significance of the proposed Articulatory Error Mapping Engine, the phoneme-level errors identified were further divided into articulatory types including the place of articulation, manner of articulation, voicing, and vowel quality. Mapping accuracy was calculated by comparing the categorization of the system with that of the experts-linguists in a sample of 300 mispronounced tokens.

TABLE V. PERFORMANCE OF THE ARTICULATORY MAPPING

Articulatory Category	Mapping Accuracy (%)
Place of Articulation	90.8
Manner of Articulation	88.5
Voicing Distinction	92.3
Vowel Quality	87.1

Table V shows the performance of the articulatory mapping module. The engine showed remarkably high mapping accuracy, and most phoneme deviations were accurately mapped into the articulatory features they were supposed to be. Such degree of readability is absent in the traditional ASR-based pronunciation devices, which can only give out scores or error indicators without attributing articulatory patterns to the errors.

E. Feature Set Contribution to Accuracy

The phoneme-level pronunciation errors in ESL students are of primary importance, to learn about the acoustic and temporal features that have the most significant influence on a successful classification. The analysis concerned trying different combinations of MFCC, spectrogram, and temporal features and how they made the difference individually and collectively. The outcomes can be used as evidence to support the effectiveness of the use of the model in providing quality, accurate, and reliable pronunciation feedback.

Fig. 9 shows how the various feature sets affect the accuracy of phoneme recognition. Accurate fused feature set was at

88.6% with slight reduction to 88.5% when using only MFCC and spectrogram, and lower with individual features: MFCC: 82.3%, spectrogram: 80.7% and temporal features: 69.4%. This shows that to a great extent, the factor that improves phoneme recognition is the combination of acoustic and time capabilities, which further increases the quality and accuracy of the feedback provided by the Articulatory Error Mapping Engine.

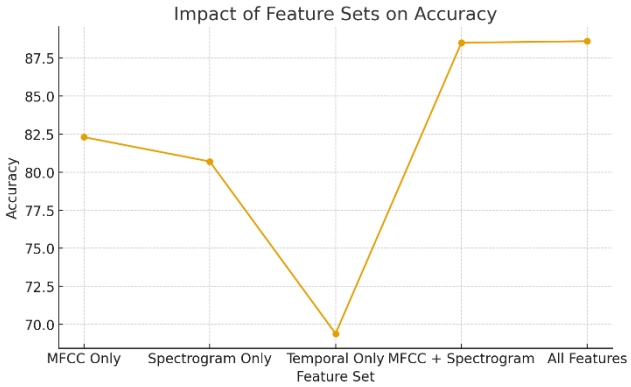

Fig. 7. Impact of feature sets on accuracy.

F. Frequency of ESL Pronunciation Errors by Type

When it comes to ESL pronunciation assessment, it is important to understand the kind of errors learners make so that it provides pointed feedback. The developed AI-based analyzer not only finds the phoneme-level inaccuracy, but also divides it into known linguistic features based on place, manner, voicing, and vowel quality.

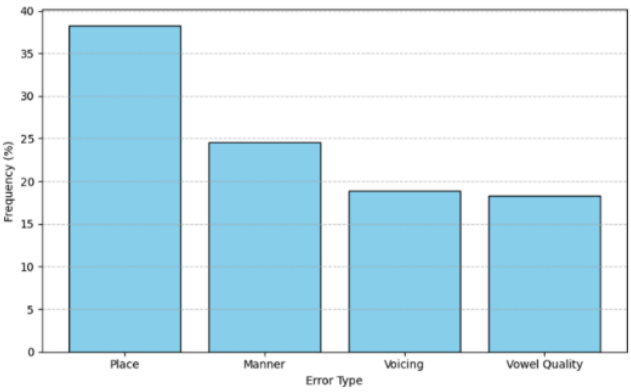

Fig. 8. Error over frequency distribution.

Fig. 10 discusses the frequency of pronunciation error types, which were detected in ESL learner speech. The most frequent mistake occurred in place of articulation (38 percent of the entire cases), which means difficulties of learners in accurate positioning of the tongue (e.g., / 0/ replaced by /t/). The substitution of affricates with stops in manner of articulation errors was followed with a rate of 24.6%. Less common yet also large were voicing and vowel quality errors, and they were about 1819%.

G. Loss vs. Epochs

Loss vs. Epoch graph indicates the decrease in error of the model during training cycles. The points will be the loss at each epoch, and this can be regarded as a measure of the efficiency of the model using the data. A constant downward trend shows the successful convergence and better the overall performance of the model.

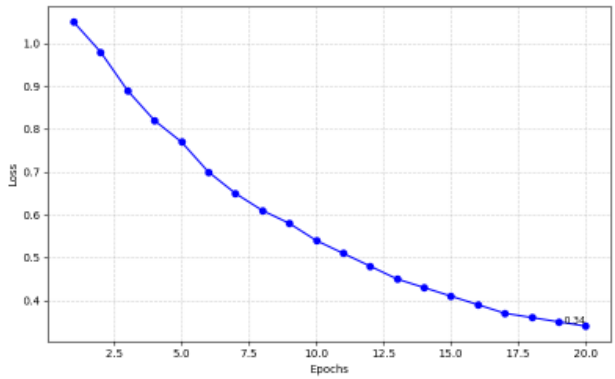

Fig. 9. Loss vs. epochs.

Fig. 11 demonstrates the value of losses at different training epochs and the tendency is to decrease. The last loss decline of 0.34 means high convergence and stable learning, that is, the CNN-BiGRU-Attention model was effective to reduce the prediction errors. This consistent trend is a confirmation that the model will provide reliable response on the training data without overfitting and that the phoneme predictions will be accurate to the downstream Articulatory Error Mapping Engine.

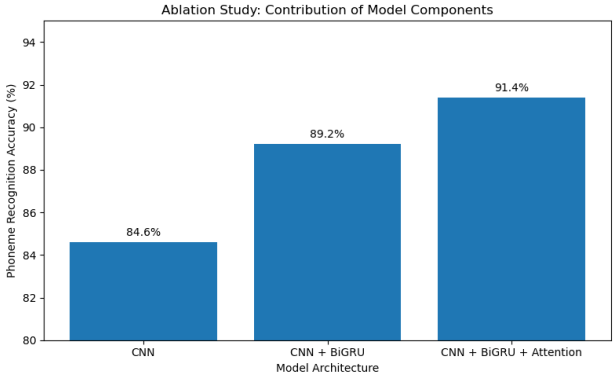

Fig. 10. Ablation study: Contribution of model components.

Fig. 12 graph shows that adding BiGRU improves temporal phoneme modeling, while the attention mechanism further increases accuracy by focusing on mispronounced phoneme segments in ESL speech.

H. Performance Comparison

A comparative analysis was carried out to determine the performance of the suggested hybrid CNN-BiGRU-Attention model in comparison with the conventional HMM-GMM models and commercial ASR-based pronunciation system. The proposed system recognized phonemes with an accuracy of 91.4 which was far much better than the ASR-based system (78.3) and the HMM-GMM base (70.5). In contrast to traditional ASR systems which focus more on the word-level decoding performance, the presented framework functions directly at the phoneme level and introduces the structured articulatory interpretation. This enhancement is not just in the classification accuracy though but also to the error granularity and

interpretability. Moreover, where large speech models based on transformers focus on large-scale representation learning, the given architecture shows that specific phoneme-level modeling with articulatory mapping can be created to attain competitive accuracy at reduced computational costs and with increased pedagogic explainability. This proves the appropriateness of the framework in assessment activities based on ESL-based pronunciation.

TABLE VI. PERFORMANCE COMPARISON

System	Phoneme Accuracy (%)
Proposed AI Hybrid System	91.4
ASR-Based Tool [16]	78.3
HMM-GMM Baseline [21]	70.5

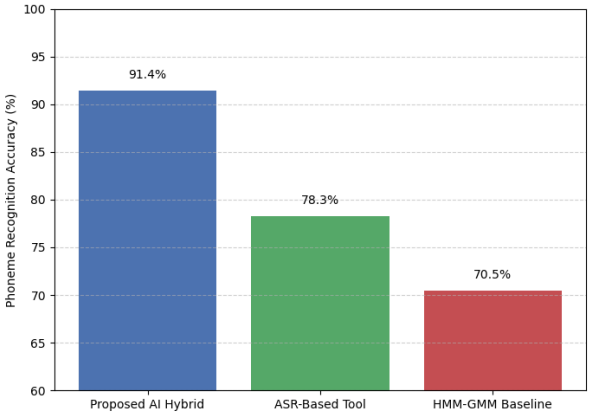

Fig. 11. Model performance comparison.

Fig.12 demonstrates the high level of the suggested AI-based hybrid system in comparison to the known pronunciation assessment tools. The proposed model has a phoneme recognition rate of 91.4 as compared to the 78.3 recorded by the ASR-based tool and the 70.5 by the HMM-GMM baseline. This demonstrates the effectiveness of that combination of CNN, Bi-GRU, and attention mechanism and a phoneme-level strategy. In every experiment, the Articulatory Error Mapping Engine leads to the difference between the proposed framework and other standard recognition systems, whereby raw recognition outputs are converted into actionable articulatory-level feedback, which has a clear pedagogical value that is absent in other standard recognition systems (ASR or HMM-based).

I. Robustness Analysis

The evaluation of robustness was conducted to evaluate the performance in the presence of realistic acoustic and speaker variations. At medium noise levels, phoneme accuracy remained relatively unchanged with the addition of background noise at 20 dB, 10 dB and 5 dB SNR levels, suggesting that spectral-temporal modeling is stable. Higher degradation was observed at very small SNR, but still the performance was higher than the traditional baselines. There was also diversity in terms of accents, gender and age. Findings showed similar phoneme recognition and articulatory classification in profiles of speakers. The comparison of pairs of phonemes that are most often confused and voiced-voiceless pairs proved that the attention layer improved discrimination. The ability of the framework to generalize was also enabled further by speaker-independent cross-validation.

J. Discussion

The experimental findings indicate that the developed hybrid CNN-BiGRU-Attention architecture is effective in both spectral and temporal features of the ESL speech, which makes it possible to recognize the phoneme level accurately. BiGRU layers make the modeling of coarticulation and timing variation greater, whereas the attention mechanism makes the distinction of mispronounced phonemes more intense. Moreover, the Articulatory Error Mapping Engine is able to map the identified phoneme errors into articulatory deviations that are interpretable to provide pedagogically relevant feedback. All in all, the framework performs well on the substitution, deletion and insertion error, much better than the traditional ASR-based and HMM-GMM systems in performance and interpretation.

V. CONCLUSION AND FUTURE WORKS

The paper introduced a hybrid AI-based phoneme recognition model to deliver valid and interpretable pronunciation evaluation of ESL students. The proposed system involved the fusion of CNN, bi-GRU, and attention with phoneme alignment and spectral-prosodic feature fusion to reach an accuracy of phoneme recognition of 91.4, compared to commercial ASR-based systems (78.3) and traditional HMM-GMM systems (70.5). The framework was quite useful in identifying errors in pronunciation with detection rates of 84, 82 and 79 percent in terms of substitutions, deletions and insertions. The suggested solution promotes intelligent training of ESL pronunciation, self-directed learning, and instructor-guided assessment; it provides a scalable, interpretable, and learner-centered method of teaching pronunciation at the phoneme level.

Further contributions to work will involve extension of the proposed framework to multilingual and tonal languages with more advanced models of prosody, stress and intonation. By incorporating transformer-based speech models with real-time mobile/web deployment, this addition will help provide extra strength, scalability and personalized pronunciation feedback to various ESL learners.

REFERENCES

[1] S. Alam, “The Conceptual Relevance of English as Lingua Franca in Non-English Speaking Countries: Revisiting History, Policies and Praxis.,” Theory & Practice in Language Studies (TPLS), vol. 13, no. 9, 2023.

[2] Z. Kupaysinova, “ENHANCING STUDENTS’LANGUAGE PROFICIENCY THROUGH INNOVATIVE INTEGRATED METHODS IN CONTEMPORARY LANGUAGE EDUCATION,” Mental Enlightenment Scientific-Methodological Journal, vol. 6, no. 03, pp. 205–217, 2025.

[3] N. Khan, A. Khan, A. Shahzadi, and others, “Students’ Pronunciation and other Languages: The Impact of L2 Interference on the Pronunciation of ESL Students,” Pakistan Journal of Society, Education and Language (PJSEL), vol. 10, no. 1, pp. 163–175, 2023.

[4] W. S. ANISSA, “The Influence of ELSA Speak Application Towards Studentsenglish Pronunciation Mastery,” Phd thesis, UIN Raden Intan Lampung, 2025.

[5] L. Yang, “Student engagement with teacher feedback in pronunciation training supported by a mobile multimedia application,” SAGE Open, vol. 12, no. 2, p. 21582440221094604, 2022.

[6] W. Sun, "The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation," Frontiers in Psychology, vol. 14, p. 1210187, 2023.

[7] T. P. Zanto et al., "Digital rhythm training improves reading fluency in children," Developmental Science, vol. 27, no. 3, p. e13473, 2024.

[8] R. El-Bialy et al., "Developing phoneme-based lip-reading sentences system for silent speech recognition," CAAI Transactions on Intelligence Technology, vol. 8, no. 1, pp. 129–138, 2023.

[9] P. Sullivan, T. Shibano, and M. Abdul-Mageed, "Improving automatic speech recognition for non-native english with transfer learning and language model decoding," in Analysis and application of natural language and speech processing, Springer, 2022, pp. 21–44.

[10] B. Maji and M. Swain, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features," Electronics, vol. 11, no. 9, p. 1328, 2022.

[11] T. S. Nguyen, T. D. T. Nguyen, N. Q. N. Hoang, and T. K. H. Do, "How AI-Powered Voice Recognition Has Supported Pronunciation Competence among EFL University Learners," Computer-Assisted Language Learning Electronic Journal, vol. 26, no. 3, pp. 64–83, 2025.

[12] S. Albaladejo Albaladejo, "Improving the English pronunciation proficiency of university students through the use of pronunciation learning strategies," Proyecto de investigación, 2022.

[13] H. P. Tiwari, "Challenges in teaching pronunciation: Secondary level English teachers' perspectives," Journal of Linguistics and Language in Education, vol. 17, no. 2, pp. 1–29, 2023.

[14] G. Sanders and A. de Bruin, "Examining the difference in error detection when listening to native and non-native speakers," Quarterly Journal of Experimental Psychology, vol. 76, no. 7, pp. 1547–1560, 2023.

[15] D. Abimanto and W. Sumarsono, "Improving English Pronunciation with AI Speech-Recognition Technology," Acitya: Journal of Teaching and Education, vol. 6, no. 1, pp. 146–156, 2024.

[16] Y. Xu, "English speech recognition and evaluation of pronunciation quality using deep learning," Mobile Information Systems, vol. 2022, no. 1, p. 7186375, 2022.

[17] N. T. Hoang, D. N. Han, and D. H. Le, "Exploring Chatbot AI in improving vocational students' English pronunciation," AsiaCALL Online Journal, vol. 14, no. 2, pp. 140–155, 2023.

[18] K. S. Raja and D. D. Sanghani, "Speech Emotion Recognition Using Machine. Learning," 2024.

[19] S. Permatasari, "Enhancing pronunciation skills in EFL students through the ELSA Speak application," Indonesian EFL Journal (IEFLJ, 2024.

[20] A. Babaeian, "Pronunciation Assessment: Traditional vs Modern Modes," Journal of Education For Sustainable Innovation, vol. 1, no. 1, pp. 61–68, 2023.

[21] A. A. Zaveri, "Hangman–A Voice-Over Game," Journal of Computing Technologies and Creative Content (JTec), vol. 8, no. 1, pp. 27–35, 2023.

[22] B. Zou, Y. Du, Z. Wang, J. Chen, and W. Zhang, "An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills," Sage Open, vol. 13, no. 3, p. 21582440231193818, 2023.

[23] Y. E. Kheir, A. Ali, and S. A. Chowdhury, "Automatic Pronunciation Assessment–A Review," arXiv preprint arXiv:2310.13974, 2023.

[24] S. Yoo and H. Ahn, "The Effects of Prosody Training with AI Chatbot on the English Pronunciation Improvement of Korean EFL Learners," 영어학, vol. 24, pp. 1300–1317, 2024.

[25] Mozilla, "Common Voice." Accessed: Jul. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/mozillaorg/common-voice