A Mispronunciation-Based Voice-Omics Representation Framework for Screening Specific Language Impairments in Children

Wei Bo
Computer Science and Engineering
University at Buffalo
Buffalo, USA
weibo@buffalo.edu

Matthew Rubino
Computer Science and Engineering
University at Buffalo
Buffalo, USA
mrrubino@buffalo.edu

Wenyao Xu
Computer Science and Engineering
University at Buffalo
Buffalo, USA
wenyaoxu@buffalo.edu

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00045

***Abstract*—This paper introduces an innovative end-to-end (E2E) framework for screening Specific Language Impairment (SLI) in children, centralizing phoneme-level mispronunciation (PLM) detection to enhance the precision and reliability. We have developed a unique voice-omics representation that translates PLM predictions into symbolic sequences, yielding significant phenotyping biomarkers that provide objective and quantifiable assessments of children's speech patterns. Through meticulous fine-tuning of the Connectionist Temporal Classification (CTC) model on the L2-ARCTIC dataset and rigorous five-fold cross-validation, our E2E models have demonstrated remarkable accuracy, with Area Under the Curve (AUC) values exceeding 0.71 and a notable recall rate of up to 71.5% on the CHILDES dataset. Our approach signifies a substantial advancement in SLI screening, leveraging cutting-edge technology to capture the complexities of spontaneous speech in children.**

***Index Terms*—SLI Screening; Phoneme-level Mispronunciation Detection; Symbolic Sequence; Phenotyping Biomarkers; Connectionist Temporal Classification Model.**

I. Introduction

Speech and language development are integral to a child's overall growth, underpinning their ability to communicate effectively and develop linguistic competence. Despite their importance, speech and language disorders or impairments are common among children, affecting 3% to 16% in the U.S., with prevalence rates fluctuating based on age and diagnostic criteria. Current evidence indicates that around 2% of children experience speech and/or language disorders severe enough to meet clinical standards, posing considerable challenges to their communication and educational development [1]. Early identification and intervention are paramount in addressing these impairments effectively, with a growing emphasis on the importance of timely screening and tailored therapeutic approaches [1], [2].

Specific Language Impairment (SLI) is a subtype of speech or language impairment with a specific focus on language difficulties that are not due to other developmental conditions. Identification of SLI in children is a multifaceted and intricate process, lacking a unified reference standard applicable to all age groups [1], [3]. Phoneme-level mispronunciation (PLM) detection has emerged as a valuable tool in the early screening of SLI. For children at risk, timely identification of mispronunciations can serve as an early indicator of potential speech or language issues, enabling prompt intervention [4]–[6]. Within the realm of automatic speech assessment, PLM detection is specifically designed for the systematic identification and categorization of deviations from standard or expected pronunciations. This technology-driven approach not only enables a detailed analysis of speech or language issues but also provides a granular understanding of individual challenges. Such detailed insights are instrumental in developing personalized speech therapy programs. Moreover, the method's precision allows for ongoing monitoring of a child's development, offering valuable feedback on the effectiveness of therapeutic interventions.

However, developing an effective screening system for SLI in children is fraught with challenges, given the unique and dynamic nature of pediatric speech. A primary challenge in constructing an end-to-end (E2E) screening system for SLI is ensuring accuracy and reliability [7]. Another significant hurdle is providing an objective and quantifiable assessment of a child's speech. Methods that are based on paralinguistic features, such as acoustic features, may fail to capture the subtle complexities of SLI [4]. Lastly, detecting fine-grained PLM in spontaneous speech is particularly challenging. Current datasets predominantly consist of repetitive single or compound words [8] or involve children reading sentences [5], which do not adequately represent the unstructured nature of spontaneous speech.

In this paper, we address these significant challenges by developing a comprehensive E2E system for early and precise screening of SLI in children, with a central focus on PLM detection. The introduced PLM detection component utilizes two methodologies - acoustic features and CTC-based automatic speech recognition (ASR) - to generate a symbolic sequence that captures objective and quantifiable phenotyping biomarkers from children's speech, providing a detailed insight into the speech patterns of children and specific SLI characteristics. A

This work is partly supported by the US National Science Foundation under DRL-2229873, OISE-2106996, and CNS-2050910.

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00045

pivotal enhancement to the system's efficiency and accessibility is the integration of OpenMP for parallel processing across CPU cores, significantly accelerating the processing speed and enabling the real-time analysis of spontaneous speech, making the tool highly suitable for naturalistic speech patterns and applicable across various settings. Moreover, rigorous validation is achieved through five-fold cross-validation to ensure the accuracy and generalizability of the entire E2E pipeline. This paper, therefore, presents a forward-thinking approach to SLI screening, merging cutting-edge technology with practical needs in pediatric speech therapy.

Our contributions are three-folds:

- PLM-based voice-omics representation framework is proposed for E2E SLI screening in children, where translating PLM prediction sequences into symbolic representation yields innovative phenotyping biomarkers for objective and quantifiable assessments.
- PLM detection, as a central component, is constructed from the perspectives of acoustic features and advanced CTC-based ASR systems.
- Performing a nuanced evaluation in a spontaneous speech scenario, employing OpenMP for accelerated processing and five-fold cross-validation for ensuring accuracy and reliability.

The remainder of this paper is organized as follows: Section II outlines the background on SLI, pediatric SLI screening and PLM detection, Section III describes the proposed E2E PLM-based voice-omics representation framework, Section IV delves into the PLM detection aspect of the E2E framework, Section V presents the benchmarking and modeling for both the E2E framework and PLM detection, and Section VI delivers an in-depth analysis of the results. Finally, the paper contains a brief discussion in Section VII, followed by the conclusion in Section VIII.

II. Literature Review

SLI affects effective communications, including speaking, listening, reading, and writing [3], [9], [10]. Early screening in children is crucial for improved outcomes [11], traditionally reliant on subjective clinical assessments by speech-language pathologists (SLPs) [1], [12]. Recently, there's been a shift towards using ASR technology and machine learning for more objective, efficient screening. These tools analyze children's speech for mispronunciations and language difficulties, enhancing screening reliability [4]–[8].

While SLI is primarily concerned with difficulties in language use, these aspects can influence, and be influenced by, speech production capabilities. PLM Detection can serve as an early indicator of underlying language processing issues, given that accurate pronunciation requires not only motor skills but also phonological processing, which is a component of language ability [13]. Also, speech and language development are highly interrelated in early childhood [14], [15]. Mispronunciations in young children can sometimes hint at broader language development issues. Thus, PLM Detection can indirectly contribute to identifying children who may require a comprehensive evaluation for SLI and capture the multifaceted nature of language impairments.

PLM Detection, vital in early SLI identification in children [16], is categorized into two main methodologies. The first involves decisive feature extraction, such as Goodness of Pronunciation (GOP) and confidence measures [17]–[19], which compare extracted acoustic features from speech against standard models to assess pronunciation quality. The second method uses Extended Recognition Networks (ERNs) to expand speech recognition search lattices, allowing a broader analysis of speech variations for improved mispronunciation detection [16], [20]. Recently, the field has evolved with the integration of E2E frameworks in ASR, particularly CTC-based methods [21], [22], which streamline PLM detection by directly learning alignments between speech and phonetic transcriptions, bypassing the need for predefined alignments or complex linguistic models, thus enhancing detection accuracy and efficiency.

The current landscape of speech and language assessments for children, while extensive, presents distinct gaps and limitations, particularly in the nuanced domain of PLM analysis and its application within speech and language pathology.

A notable limitation in current methodologies in speech and language assessments, such as those by Black et al. [23] and Duchateau et al. [24], is the tendency to detect word-level disfluencies for assessments. This approach overlooks the complexities at the phoneme level, which are crucial for a comprehensive understanding and addressing of subtle speech impairments. Therefore, a shift towards more detailed, phoneme-specific analysis is essential for accurately evaluating and intervening in children's speech or language development.

In the realm of PLM detection for children's language assessments, existing research often relies on datasets with limited scope, mainly focusing on single-word pronunciations or sentence readings. Studies like Yilmaz et al. [25], Proença et al. [5], and Hair et al. [8] have primarily used tasks involving word and sentence reading to detect pronunciation errors. However, these datasets fall short of capturing the complexities of spontaneous speech, which more accurately reflects children's natural speech patterns. To address this, our approach incorporates the CHILDES Clinical English ENNI Corpus [26], [27], which utilizes narrative elicitation from storybooks or picture sequences, offering a more holistic and realistic analysis of children's speech or language capabilities.

A further limitation lies in the absence of quantifiable and objective measurement of identification. For example, Shahin et al. [4] used paralinguistic features in the acoustic area to directly construct a screener, and Proença et al. [5] extracted features with the consideration of variants and coarticulation rules. However, those features lack insights for SLPs. Essential information, such as the identification of mispronounced phonemes or frequency of errors, is often missing. This level of detail is essential for SLPs to effectively tailor therapy plans and provide focused intervention, addressing the unique needs of each child.

We recognize the absence of a unified standard in SLI

screening and the consequential difficulty in establishing a direct performance comparison. By identifying several limitations in current methodologies, we propose an objective, quantifiable, and phoneme-specific analysis on natural speech for accurately evaluating and intervening in children's speech or language development. However, the innovative nature of our method challenges direct comparisons with traditional tools, which fail to capture the naturalistic speech patterns we prioritize. The differences in methodologies and datasets further diminish the relevance of direct comparisons.

III. E2E Framework Overview

We propose an innovative E2E pipeline, which is a mispronunciation-based voice-omics representation framework for screening SLI in children. This framework is meticulously crafted to offer automated, precise, reliable, and early screening tailored for pediatric SLI, leveraging PLM detection to identify distinct phenotyping biomarkers in children's spontaneous speech for enhanced accuracy. As Figure 1 shows, the framework consists of three integral components:

A. Audio Preprocessing

In the preliminary phase of our framework, raw audio recordings that capture children engaging in storytelling activities, prompted by a series of stimuli images, are processed. The fundamental objective in this stage is to refine these recordings to focus exclusively on the children's speech. To achieve this, we employ advanced speaker diarization techniques, utilizing the 'pyannote.audio' speaker diarization pipeline, specifically version 2.1 [28], [29]. This technology is adept at discerning and segregating different speakers within the audio. By applying this pipeline, we efficiently isolate the children's speech from the overall audio mix, which includes the investigator's speech and other extraneous sounds [26], [27]. This step also involves the removal of all silent intervals, including both silence within the children's speech and those resulting from speaker transitions, thereby eliminating any pauses that do not contribute to the speech content analysis. Finally, speech segments are concatenated to create an uninterrupted audio stream for each child, which provides a clean and focused dataset for the subsequent stages of PLM detection and SLI screening. This preprocessing is crucial for the accuracy and reliability of our analysis, as it ensures that our system evaluates only the relevant speech data, thus enhancing the overall effectiveness of the screening tool.

B. PLM Detection

The central component of our framework is PLM detection, which we have segmented into two distinct approaches: Acoustic-Based Detection (ABD) and Transcription-Based Detection (TBD). This section offers a preliminary overview, with comprehensive details to be presented in Section IV.

In our Acoustic-Based Detection (ABD) methodology, we first transform preprocessed audio into word-level text using orthographic ASR, followed by forced alignment to obtain phoneme-level timestamps, converting the audio into phoneme segments. We then extract specific acoustic features for each phoneme and develop a binary classifier to create a voice-omics representation. Alternatively, the Transcription-Based Detection (TBD) method utilizes a CTC-based phoneme ASR technique for direct transcription of audio into phoneme-level text and phoneme segmentation. The phoneme segments are converted into the CMU ARPABET format [30], upon which we construct a comparative model to derive the voice-omics representation.

In both methods, each phoneme in the audio recordings receives a binary label, either 'C' for correct pronunciation or 'E' for errors, from the binary classifier or comparative model. This process creates a mispronunciation detection (MD)-based phenotyping sequence, offering a detailed view of the child's phonemic accuracy.

C. SLI Screener

The right side of our framework delineates the process employed for screening SLI in children. Building upon the symbolic sequence from PLM detection, a variety of phenotyping biomarkers are extracted from several analytical perspectives, including density, run-length encoding, and sequence complexity. Such a multifaceted approach allows us to capture a comprehensive profile of each child's speech pattern, crucial for an accurate SLI assessment. Finally, these phenotyping biomarkers are employed to construct and fine-tune Support Vector Machine (SVM) and Random Forest (RF) models. This is referred to as Sequence-Based Screening (SBS). These classifiers are designed to evaluate the likelihood of SLI in children based on the analyzed speech patterns, which ensures that the assessment is grounded in objective, quantifiable speech characteristics. These two classical machine learning models were selected because they have been the de facto standard for the classification task, and it would be less challenging for general users to understand. Also, classical models often serve as good benchmarks. Consequently, this portion of the framework is integral to achieving a reliable and effective tool for early SLI detection in pediatric populations.

IV. PLM Detection

This section will elaborate on the specific implementation details of PLM detection, the core of our E2E framework.

A. Acoustic-Based Detection (ABD)

Figure 2 shows the detailed view of ABD.

1) Phoneme-based Segmentation:

Orthographic ASR: The preprocessed audio signal undergoes intricate processing steps using Whisper-Medium, an advanced ASR system developed by OpenAI [31]. This system is adept at handling extensive audio data, segmenting the entire audio stream into discrete 30-second chunks for analysis. Each segment is then meticulously transcribed into orthographic text, ensuring its accuracy and linguistic correctness.

Forced Alignment: Next, Montreal Forced Aligner (MFA), a linguistic tool designed for aligning speech audio with its corresponding text transcription [32], is applied to get precise

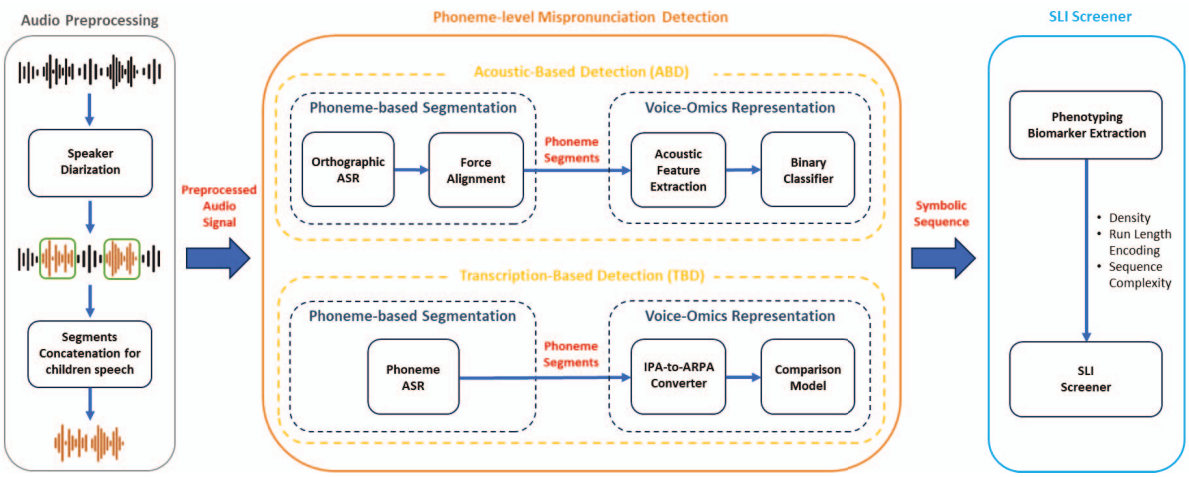

Fig. 1. A mispronunciation-based voice-omics representation framework for screening SLI in children

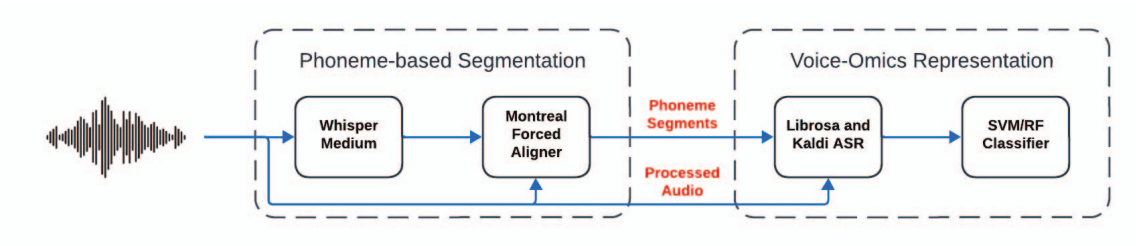

Fig. 2. A detailed view of the acoustic-based detection (ABD) framework

phoneme-level alignment. The MFA processes these inputs by extracting features from the audio and utilizing a phonetic dictionary to map words to their phonetic representations. This process yields timestamps for each phoneme, effectively segmenting the audio data into distinct phoneme units.

2) Voice-Omics Representation:

Acoustic Feature Extraction: Then, Librosa [33] and Kaldi [34] are used to extract acoustic features for each phoneme using the phone-level timestamps produced by the forced alignment. Kaldi is used to compute GOP, and Librosa is used to compute the remaining acoustic features described in Section V-B2.

Binary Classifier: Finally, the acoustic features for each phoneme are processed through sophisticated machine learning models, including SVM and RF. These models are employed to perform binary PLM classification to determine whether each phoneme is pronounced correctly. These two models were selected for the same reasons stated in Section III-C

B. Transcription-Based Detection (TBD)

Figure 3 shows the detailed view of TBD.

1) Phoneme-based Segmentation:

Phoneme ASR: The preprocessed audio signal undergoes the phonetic transcription process using XLSR-Wav2Vec2, a CTC-based ASR created by Facebook [35]. The model was fine-tuned for the phone-level transcription task in this study, achieving phone error rates (PER) comparable to state-of-the-art [36]. The output International Phonetic Alphabet (IPA) phones are grouped by space-separated words, which is critical for the subsequent steps in the pipeline.

2) Voice-Omics Representation:

IPA-to-ARPA Converter: Next, Gruut IPA, a tool for manipulating IPA pronunciations [37], is used to convert the IPA transcriptions created in the previous step to CMU ARPABET. Because each IPA symbol maps to one symbol in CMU ARPANET (disregarding stress) [30], this process resembles a traditional mapping operation. This step is necessary to ensure the phonetic transcription uses the same phonetic alphabet as the comparison model.

Comparison Model: Finally, each word in the phonetic transcription is isolated and used to query the CMU Pronouncing Dictionary, which contains pronunciations for over 134,000 English words in CMU ARPABET [30]. The query is performed by doing a linear scan, parallelized with OpenMP, over the entire dictionary to locate the word with the closest pronunciation by Levenshtein distance. An alignment is then computed between the actual pronunciation and the target's closest pronunciation using the Needleman-Wunsch algorithm [38], labeling each phone as correct if it matches the target in the alignment and erroneous otherwise.

More formally, Levenshtein distance describes the minimum number of operations required to convert a source string to a target string, where valid operations are insertion, removal, and substitution. Levenshtein distance $\mathrm{LD}(s,t)$ can be described by the following recursive definition, where s is the source string of length n, t is the target string of length m, and x_i indicates the character at position i in string x:

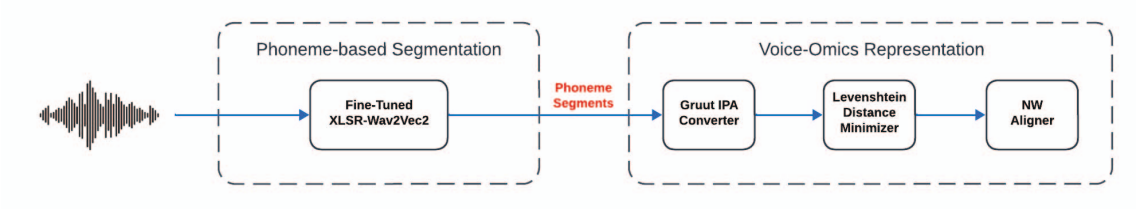

Fig. 3. A detailed view of the transcription-based detection (TBD) framework

$$\mathrm{LD}(s,t) = \begin{cases} |t| & \text{if } |s| = 0 \\ |s| & \text{if } |t| = 0 \\ \mathrm{LD}(s_{2:n}, t_{2:m}) & \text{if } s_1 = t_1 \\ 1 + \min \begin{cases} \mathrm{LD}(s_{2:n}, t) \\ \mathrm{LD}(s, t_{2:m}) \\ \mathrm{LD}(s_{2:n}, t_{2:m}) \end{cases} & \text{otherwise} \end{cases} \quad (1)$$

Then, given a word s from the source ARPABET transcription, we compute the target pronunciation t as follows, where D is the set of all pronunciation strings in the CMU Pronouncing Dictionary:

$$t = \underset{t^* \in D}{\operatorname{argmin}} \, \mathrm{LD}(s, t^*). \quad (2)$$

Finally, we perform the Needleman-Wunsch algorithm to compute an alignment between s and t. The resulting aligned strings s' and t' have equal length l, and each character is either in the set of ARPABET phones or is the blank character. Then, we label each phone p_i as follows, where $1 \leq i \leq l$:

$$p_i = \begin{cases} C & s'_i = t'_i \\ E & \text{otherwise} \end{cases}. \quad (3)$$

V. Benchmarking and Modeling

This section describes the configuration for all of our benchmarking and modeling. The first part focuses on our E2E SLI screening framework, and the second part is for our core PLM detection component. We introduce datasets, feature spaces, model configurations, and evaluation metrics for each part.

A. E2E framework

1) Datasets: CHILDES Clinical English ENNI Corpus [26], [27] has been used for benchmarking our E2E SLI screening framework. It encompasses a diverse range of narrative data from English-speaking children. This includes both typically developing (TD) children and those with SLI, which is compiled from a cohort of children aged 4 to 9, comprising 77 participants with SLI and 300 TD participants. This rich corpus consists of narrative samples elicited from children through a series of picture stimuli, specifically designed to encourage storytelling, which provides a natural context for studying children's narrative skills and makes this corpus crucial for understanding various aspects of language acquisition and identifying language disorders in children.

For our specific research purposes, we focused on the subset of the dataset that included audio recordings. This decision was necessitated by the fact that some children in the corpus were represented only through transcripts without corresponding audio. Consequently, our analysis incorporated data from 67 children diagnosed with SLI and 288 TD children, forming the basis for our E2E pipeline and the model training and testing for the SLI Screener component. Each selected speaker in this subset contributed one audio recording. These recordings, spanning several minutes, encompass not only the speech of the child but also the interactions with the investigator.

2) Feature Space - Phenotyping Biomarkers: Phenotyping biomarkers are extracted from the symbolic sequence generated by PLM detection stage, which is a string containing the characters 'C' and 'E', where 'C' indicates a correctly pronounced phoneme and 'E' indicates the mispronounced phoneme. These phenotyping biomarkers are designed to provide objective and quantifiable speech characteristics for effectively screening SLI in children.

The detailed information and equations for each phenotyping biomarker are shown below, where S represents a length n string of 'C' and 'E' characters, and S_i is the character at position i such that $1 \leq i \leq n$. Note that functions $c : S \to \{S\}$ and $e : S \to \{S\}$ are defined as taking a string S and returning a list of all contiguous 'C' and 'E' subsequences respectively. Additionally, should the expression for any metric be undefined (e.g. ACE when there are no errors), it is treated as 0. Finally, we adopt the Iverson bracket notation [39] which is defined as follows:

$$[P] = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

- Mispronunciation Density (MPD): MPD describes the relative frequency of mispronunciations in a section of speech. It is designed from the density perspective, where a high MPD value indicates high anomalies or variations during speech. We expect children with SLI to have a positive correlation value on this biomarker.

$$MPD = \frac{1}{n} \sum_{i=1}^{n} [S_i = E] \quad (5)$$

- Normalized Transition Count (NTC): NTC describes how frequently the speaker transitions between mispronounced and correctly pronounced phones. NTC is designed from the sequence complexity perspective, where this biomarker indicates the information flow or sequence

dynamics. A high NTC value means more frequent changes or shifts, and we expect children with SLI to have a positive correlation value on this biomarker.

$$NTC = \frac{1}{n}\sum_{i=2}^{n}[S_i \neq S_{i-1}] \tag{6}$$

- Average Common Correct (ACC): ACC describes the average number of successive phones the speaker pronounces correctly. It's designed from the sequence complexity perspective. We expect to see a negative correlation between this biomarker and children with SLI because we believe children with SLI will have less correctly pronounced phones (greater MPD) and more transitions (greater NTC) in a section of the speech.

$$ACC = \frac{1}{|c(S)|}\sum_{i=1}^{n}[S_i = C] \tag{7}$$

- Average Common Error (ACE): ACE describes the average number of successive phones the speaker pronounces incorrectly. Similar to ACC, ACE is also designed from a sequence complexity perspective. It's challenging to hypothesize about the correlation between ACE and children with SLI because they are expected to have more incorrectly pronounced phones and transitions in a section of speech.

$$ACE = \frac{1}{|e(S)|}\sum_{i=1}^{n}[S_i = E] \tag{8}$$

- Longest Common Correct (LCC): LCC describes the maximum number of successive phones the speaker pronounces correctly. LCC is designed from the run length encoding perspective, where a higher LCC value potentially indicates fewer errors in a section of speech. We expect to see a negative correlation between LCC and SLI.

$$LCC = \max_{S^* \in c(S)} |S^*| \tag{9}$$

- Longest Common Error (LCE): LCE describes the maximum number of successive phones the speaker pronounces incorrectly. Similar to LCC, LCE may indicate continuous error presence and a high LCE can reveal strong connections or associations of the errors. We expect to see a positive correlation between LCE and SLI.

$$LCE = \max_{S^* \in e(S)} |S^*| \tag{10}$$

3) Model Configurations for SLI Screener: When selecting model configurations for the SLI screener, we applied five-fold cross-validation and selected the hyperparameters that maximized AUC. This helps us ensure the accuracy and generalizability of the models across different subsets of data. For the ABD SBS SVM, the best configuration was a linear kernel with a C of 0.01. For the TBD SBS SVM, the best configuration was a linear kernel with a C of 100. For the ABD SBS RF, we used 40 estimators with a maximum depth of 5. Finally, for the TBD SBS RF, we used 5 estimators with a maximum depth of 5.

4) Evaluation Metrics for SLI Screener: SLI screeners were evaluated using AUC, accuracy, precision, recall, and F1. A sample with SLI was considered positive, and a sample with TD was considered negative. For this task, we considered the most important metrics to be AUC and recall because AUC provides a global view of the model's performance across all thresholds, and recall gives insight into how effectively the model avoids false negatives. High values for these metrics reduce the probability of missing SLI-positive children with the screener, which is one of our main aims.

B. PLM Detection

1) Datasets: The L2-ARCTIC speech corpus, designed for voice and accent conversion research as well as mispronunciation detection, is instrumental in our PLM detection stage [40]. Comprising high-quality audio from 24 non-native English speakers across six languages (Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese) and providing orthographic and forced-aligned phonetic transcriptions, the corpus also includes 150 manually annotated utterances per speaker. These annotations, identifying common mispronunciation errors like substitutions, deletions, and insertions, enhance the dataset's utility for pronunciation training and speech recognition accuracy. We leverage this detailed subset for fine-tuning, training, and testing in PLM detection, treating phonemes marked as errors as mispronunciations.

It is important to mention that the dataset presented an imbalance with a majority of phonemes labeled as correctly pronounced. To counteract this and ensure effective model training and evaluation, we employed random downsampling to attain a balanced dataset, with an equal count of 8017 correctly and mispronounced phonemes. For the training phase, we allocated 6413 correctly pronounced and 6414 mispronounced phonemes, reserving the rest for testing. Furthermore, the extensive quantity of phonemes presented a computational challenge for conventional classifiers like SVM, potentially impeding their ability to converge swiftly. Through selection and equalization of phoneme quantities, we optimized the classifiers' performance, thereby enhancing the precision and dependability of our PLM detection results.

2) Feature Space - Acoustic Features: Several acoustic features are extracted for each phoneme by acquiring the phone-level timestamps (either from labels or MFA), slicing the audio at each phoneme, and computing the corresponding features on the audio signal. These acoustic features are used for the ABD method in this PLM detection component. The detailed information for each acoustic feature is shown below.

- First three formants (F1, F2, and F3): Formants are resonant frequencies of the vocal tract, and they play a significant role in characterizing how vowels sound.
- The first 13 Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are the representations of the short-term power spectrum of a sound, which are essential

in speech recognition, speaker identification, and audio classification.
- Spectral contrast (7 bands): This feature analyzes the amplitude differences between peaks and valleys in an audio spectrum across seven distinct frequency bands.
- Spectral bandwidth: It quantifies the range of frequencies encompassing the majority of a sound signal's energy, highlighting the frequency spread within a sound.
- GOP score: GOP score offers a numerical assessment of pronunciation quality, comparing how closely a sound or phoneme matches standard or native pronunciation. It's computed by taking the log of the posterior probability of phone p given evidence $O^{(p)}$ and then normalizing the result.

$$\begin{aligned} \mathrm{GOP}(p) &\equiv \left|\log(P(p|O^{(p)}))\right| / \mathrm{NF}(p) \\ &\approx \left|\log\left(\frac{P(O^{(p)}|p)}{\max_{q\in Q} P(O^{(p)}|q)}\right)\right| / \mathrm{NF}(p) \end{aligned} \tag{11}$$

3) Model Configurations for PLM Detection:

ASR Models: An XLSR-Wav2Vec2 model was used for phonetic transcription in TBD. This CTC model was selected for its effectiveness in phoneme-level mispronunciation detection. It is preferred because it directly learns alignments between speech and phonetic transcriptions, bypassing the need for predefined alignments or complex linguistic models. This improves detection accuracy and efficiency. Also, this model was fine-tuned on the L2-ARCTIC dataset to improve recognition accuracy for phoneme-level variations and mispronunciations in children's speech. The decision to fine-tune a model on L2-ARCTIC was motivated by the fact that the TBD pipeline requires transcriptions of exactly what a speaker said, rather than their intent. Existing XLSR-Wav2Vec2 models are trained on phonetic labels that do not include mispronunciation information [41]. Because we needed to capture mispronunciations in the transcription for TBD to work, a model tuned on L2-ARCTIC was more theoretically sound. To fine-tune the model, the manually annotated utterances from L2-ARCTIC were divided into a 90/10 train/test split, with 3224 training utterances and 359 testing utterances. The training utterances were then used to fine-tune the base XLSR-Wav2Vec2 model, and the testing utterances were used to evaluate the model.

Binary Classifier Models: Similar to the SLI screener classifiers, the SVM and RF model configurations for PLM detection were selected by applying a five-fold cross-validated grid search across their common hyperparameters and choosing the configuration that maximized AUC on their respective datasets to ensure accuracy and generalizability. For the SVM, this was the radial basis function kernel with C of 10 and an automatically scaled gamma. For the RF, this was 200 estimators with an infinite depth. Finally, all features were standardized to unit mean and variance before training and evaluation.

4) Evaluation Metrics for PLM Detection: For PLM detection, ABD and TBD were evaluated in a slightly different manner, which was necessitated by differences in the ways each approach generated predictions.

For ABD, the evaluation was straightforward, using the same metrics as the SLI screeners. This included AUC, accuracy, precision, recall, and F1, defined in Section V-A4. The positive class label indicated a correctly pronounced phone, and the negative class label indicated a mispronounced phone.

For TBD, the evaluation was broken into two components. The first component was the transcription performance of the ASR models. This was measured in terms of the word error rate (WER) and phone error rate (PER) between predicted sequences and actual sequences.

The second component was the detection performance of the comparison model. This was measured using the same metrics as ABD with the exception of AUC, which could not be obtained for TBD since it did not use a variable decision boundary. Additionally, unlike ABD, the predicted CE sequence for TBD p could have a different length than the labeled CE sequence y. Thus, to evaluate TBD, we aligned sequences p and y using the Needleman-Wunsch algorithm to produce sequences p' and y' with equal length l. We then used all 1786 pairs from the testing dataset where $y'_i = E$ and randomly sampled 1786 pairs where $y'_i = C$, applying the following rule to each pair of p'_i and y'_i where $1 \leq i \leq l$:

$$\begin{cases} TP & y'_i = C = p'_i \\ TN & y'_i = E = p'_i \\ FP & y'_i = E \neq p'_i \\ FN & y'_i = C \neq p'_i \end{cases} \tag{12}$$

As a final note, some papers break down true rejections (in this case, TN) for PLM detection into subcategories 'correct diagnosis' and 'diagnosis error' [16]. That is not done in this paper because phones are only labeled 'C' or 'E', meaning errors are not diagnosed in this work.

VI. Experimental Results and Analysis

In this section, we present results for both the E2E SLI screening framework and PLM detection. For the E2E framework, we focus on highlighting its performance using phenotyping biomarkers compared with a baseline, followed by a statistical analysis of these biomarkers. In PLM detection, we evaluate the binary classification performance of ABD and assess the transcription accuracy of the fine-tuned ASR model and the performance of the comparison model in TBD.

A. E2E Framework

1) SLI Screener Performance: To establish a baseline for SLI screening, we first implemented acoustic-based screening (ABS). Differing from sequence-based screening (SBS), ABS predicts for each pronounced phone whether or not it was produced by a speaker with SLI and computes the final label of a speaker based on a majority vote over a set of phones. Afterwards, we developed E2E SLI screening models using SVM and RF techniques, utilizing the phenotyping biomarkers detailed in Section V-A2. These models were built on the

symbolic sequences generated respectively by the RF classifier of ABD and TBD. This choice was informed by our PLM performance analysis, which will be elaborated in Section VI-B. This structured approach ensures a comprehensive and methodical evaluation of SLI screening methodologies.

The performance for all the screeners is shown in Table I. Two baselines (ABS SVM and ABS RF) have distinctive performances. Specifically, the ABS RF screener has better AUC, accuracy, and precision, while the ABS SVM has better recall and F1 score. For E2E SLI screeners, our model demonstrates robust and consistent performance in terms of the AUC, with the ABD SBS SVM model achieving the highest AUC at 0.725. Furthermore, the SVM models applied to both ABD SBS and TBD SBS have exhibited superior recall rates. Notably, the TBD SBS SVM model has attained the highest recall, reaching 0.715.

TABLE I
SLI SCREENER PERFORMANCE SUMMARY

	AUC	Accuracy	Precision	Recall	F1
ABS SVM	0.669	0.670	0.305	0.567	0.396
ABS RF	0.701	**0.792**	**0.426**	0.266	0.304
ABD SBS SVM	**0.725**	0.662	0.319	0.674	**0.430**
ABD SBS RF	0.716	0.707	0.326	0.466	0.375
TBD SBS SVM	0.710	0.608	0.286	**0.715**	0.408
TBD SBS RF	0.710	0.746	0.345	0.357	0.348

We also show the five-fold cross-validated ROC curves for E2E ABD SBS and E2E TBD SBS in Figure 4 and Figure 5 respectively. The AUC values for the SVM models seem more variable across the folds, which might indicate a sensitivity to the data distribution in each fold. The RF models appear to be more robust with less variation in AUC, especially in TBD SBS settings. The variance in AUC scores across different folds suggests that model performance may be influenced by the particular characteristics of the data in each fold, which could include the distribution of SLI and non-SLI cases or the complexity of the speech samples.

One thing to note is that while the ABS RF model shows better performance on some metrics, these metrics are less important for the SLI screening task. For SLI screening, it is most important to catch all SLI-positive children (i.e., avoid false negatives). Since the ABS RF model has the worst recall, it is clearly a poor choice for this task. In addition, our proposed SBS method improves model interpretability and explainability, since sequence features like LCE are more understandable and quantifiable than acoustic features like 13 MFCCs. Therefore, our SBS method surpasses the baseline in key metrics and enhances understandability, making it a highly valuable approach.

Furthermore, for SBS models, the ABD SBS SVM stands out with the highest AUC and F1 score, suggesting it's the most effective for SLI screening, balancing SLI detection (recall), and minimizing false positives (precision). Despite lower precision and accuracy, the TBD SBS SVM exhibits the highest recall, making it superior in identifying positive cases

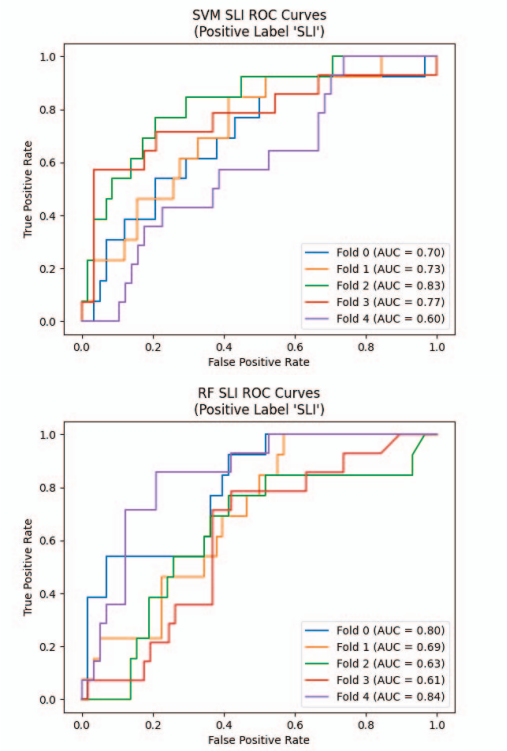

Fig. 4. Five-fold cross-validated ROC curves in E2E ABD SBS setting

of SLI, a critical factor in contexts where missing a diagnosis is costly.

2) Phenotyping Biomarkers Analysis: Next, we performed statistical analysis on the phenotyping biomarkers described in Section V-A2, where sequences were generated using RF ABD and TBD. We performed this analysis by calculating the Point-Biserial Correlation Coefficient (r value) for each biomarker with respect to SLI status and chose p = .050 as a cutoff for statistical significance. Point-Biserial Testing is a special case of the Pearson correlation coefficient and measures the strength and direction of the association between a continuous variable and a binary categorical variable. The correlation coefficient can range from -1 to 1, where values close to -1 or 1 indicate a strong relationship, and values near 0 indicate a weak relationship. The r and p values for each feature are shown in Table II and Table III.

TABLE II
ABD PHENOTYPING BIOMARKER CORRELATION

	MPD	NTC	LCC	ACC	LCE	ACE
r value	0.312	-0.085	-0.199	-0.297	0.149	0.300
p value	**1.87E-9**	0.109	**1.64E-4**	**1.10E-8**	**4.84E-3**	**1.90E-9**

TABLE III
TBD PHENOTYPING BIOMARKER CORRELATION

	MPD	NTC	LCC	ACC	LCE	ACE
r value	0.298	0.306	-0.267	0.104	0.104	0.023
p value	**1.00E-8**	**3.81E-9**	**3.17E-7**	**9.91E-8**	**0.050**	0.662

The results show that most phenotyping biomarkers (MPD, LCC, ACC, LCE) are statistically significant for both ABD and

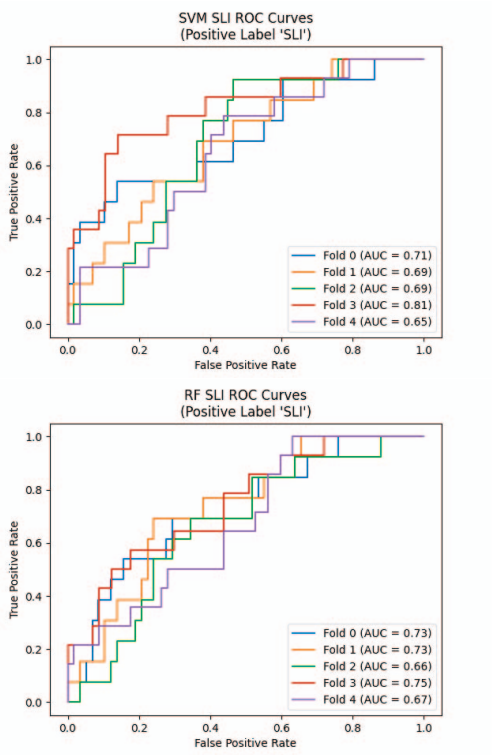

Fig. 5. Five-fold cross-validated ROC curves in E2E TBD SBS setting

TBD. In addition, our expectations for correlation direction have been verified for MPD, ACC, and LCE, indicating that the phenotyping biomarkers we constructed from sequence density and run-length encoding are effective and reliable. The three biomarkers built from the perspective of sequence complexity show some fluctuations, such as high p value or inconsistency in the direction of coefficients on ABD and TBD, due to the complexity of the speech.

B. PLM Detection

We conducted independent evaluations of our ABD and TBD methods, thereby gaining a thorough insight into the performance of this pivotal component within the E2E framework.

1) ABD: The results for the ABD approach, which used the features from Section V-B2, are shown in Table IV, where the RF model outperformed the SVM model on all key metrics.

TABLE IV
ABD PERFORMANCE

	AUC	Accuracy	Precision	Recall	F1
SVM	0.663	0.619	0.617	0.628	0.623
RF	**0.697**	**0.644**	**0.632**	**0.689**	**0.659**

2) TBD: For TBD, we first conducted error analysis where we identified the rate of inaccuracies introduced by the ASR system. It involved comparing ASR-generated transcriptions with manually verified transcriptions to quantify the misinterpretation rate. We fine-tuned XLSR-Wav2Vec2 on the L2-ARCTIC manually annotated subset to alleviate the possibility of the ASR model misinterpreting incorrect pronunciations as errors in transcription. The evaluation achieved a WER of 42.5% and a PER of 12.8%, comparable to state-of-the-art on the TIMIT dataset [36]. The WER suffered due to the absence of a language model, which typically improves WER through contextual prediction. We omitted this integration to avoid masking mispronunciation data, prioritizing clarity in phonetic analysis.

In TBD performance evaluation, the model achieved a detection accuracy of 0.604 and precision of 0.846 but showed lower recall and F1 scores of 0.255 and 0.392, respectively. The suboptimal results may be linked to the large pronunciation lexicon. The CMU Pronunciation lexicon, containing many rarely used words, especially in children's speech, can cause mismatches with mispronounced words, leading to a skewed 'C' label prediction and increased false positives. A potential solution is to condense the lexicon to words more frequently used by children.

To speed up TBD, we parallelized the lexicon search over multiple CPU cores using OpenMP and benchmarked the process on an Intel Core i5-12600K Processor. The mean runtime for each lexicon search was calculated for child #413 (SLI) in the CHILDES dataset, as benchmarking the entire dataset was time-intensive. Results are detailed in Table V.

TABLE V
LEXICON SEARCH PERFORMANCE

	Mean Runtime (s)	Mean Speedup
Python	1.215	1.0
C	0.051	22.1
C + OpenMP	**0.035**	**34.7**

VII. DISCUSSION AND FUTURE PLAN

A. Discussion and Insights

In this study, we developed an E2E screening system for SLI in children, with a focus on PLM detection. This automated tool excels in extracting objective, quantifiable phenotyping biomarkers from children's speech, particularly emphasizing the analysis of detailed speech patterns in spontaneous speech. This method provides a nuanced and comprehensive approach to SLI screening in pediatric populations.

1) PLM-based Voice-Omic Representation Framework:

Innovative Phenotyping Biomarkers: Our study introduces novel phenotyping biomarkers for children's speech, derived from the symbolic sequences in PLM detection. Most of the biomarkers have shown statistical significance with p-values below 0.05, affirming their reliability and validating the effectiveness of our E2E framework in pinpointing SLI characteristics.

Core PLM Detection Component: The core of our system, PLM detection, was built from distinct ABD and TBD methodologies. This dual approach, incorporating both acoustic feature analysis and transcription, enables thorough speech pattern assessment. Utilizing a cutting-edge CTC-based model, our framework proficiently processes and analyzes speech at a phoneme-specific level, facilitating precise SLI screening in children.

2) Accurate, Reliable and Generalized E2E SLI Screening Pipeline:

Performance of E2E Models: Our four E2E models for SLI screening demonstrated robust performance, with each achieving an AUC value over 0.71, indicating strong accuracy and predictive power. Notably, the TBD SBS SVM model exhibited a high recall rate of 71.5%, crucial for effectively identifying SLI cases.

Fine-Tuning of the CTC Model: Fine-tuning the CTC model on the L2-ARCTIC dataset was vital to meet the E2E pipeline's needs, which necessitates accurate transcriptions of what was spoken, as opposed to speaker intent. This step inherently introduces a level of linguistic diversity into the model as the L2-ARCTIC dataset comprises speakers across six different native languages, making it more robust against variations in pronunciation that could be attributed to cultural or linguistic backgrounds. It also significantly enhanced the model's capability to accurately reflect the speech of children with SLI, especially in spontaneous speech.

Reliability through Cross-Validation: We employed five-fold cross-validation across all experiments to guarantee the robustness and wider applicability of our findings, which can be seen as a step toward generalization. Each fold likely contains a variety of speech patterns, which helps ensure reliable performance metrics and that the model does not overfit to a particular subset of data and can generalize across different speech types or to diverse populations.

Evaluated in Spontaneous Speech: The CHILDES dataset, which was used for benchmarking, includes narrative samples elicited from children through picture stimuli, encouraging natural speech. This real-world application of the model to spontaneous narrative speech can enhance its ability to generalize across different cultural and linguistic backgrounds, as storytelling often reveals deep-seated linguistic structures.

B. Future Plan

In our future work on E2E SLI screening for children, we plan to implement two key strategies:

Enhanced Phenotyping Biomarkers Based on Sequence Analysis: Our current work has laid a foundation by extracting phenotyping biomarkers from the overall view of the symbolic sequence. Moving forward, we aim to delve deeper and explore these biomarkers from the segmented perspective of correct and error phonemes, which may uncover nuanced and neglected speech patterns indicative of SLI in children. We will also investigate the distribution of correct and error phonemes, their trends, and cyclical patterns within speech data. This approach is anticipated to reveal intricate speech patterns that are characteristic of SLI, thereby enriching our phenotyping palette and potentially enhancing the diagnostic precision of our framework.

Innovative Sequence Generation Strategies: In our existing framework, the sequences are generated based on the CE symbolic sequence without accounting for temporal dynamics. We intend to evolve our sequence generation strategy by incorporating a time-series dimension. This forthcoming models will incorporate actual time information, thereby transitioning from a solely symbolic to a time-sequenced analysis of speech. By embedding time-series data, we anticipate capturing the temporal dynamics of speech or language formation and usage in SLI children.

Reliability Assessments and Enhancements of ASR: The inherent challenge arises when the ASR model potentially misinterprets incorrect pronunciations as errors in transcription, leading to inaccurately labeled data. We intend to assess how these transcription inaccuracies affect our framework's ability to accurately identify SLI firstly. This may involve re-evaluating our dataset with manually verified labels to measure any significant changes in the framework's performance metrics. Also, to mitigate the impact of ASR inaccuracies, we would like to explore the feasibility of customizing and retraining the ASR model on a dataset more representative of our target demographic (children's speech, including common mispronunciations and speech impairments), to reduce the model's misinterpretation of incorrect pronunciations.

Generalization Enhancement to Cultural and Linguistic Diversity: We plan to include more varied datasets that capture a broader spectrum of cultural and linguistic backgrounds. This expansion aims to enhance the model's ability to learn and generalize speech patterns across diverse populations. And future iterations of the model could focus on incorporating culturally sensitive approaches (like adaptive layers) that account for variations in language use and expressions. The phenotyping biomarkers that were derived from the symbolic sequences in PLM detection could be further analyzed to determine if certain patterns are universally indicative of SLI across cultures and languages, or if new biomarkers need to be developed to account for cultural and linguistic diversity.

VIII. Conclusion

In conclusion, our E2E mispronunciation-based voice-omics representation framework represents a significant leap forward in the early screening of SLI in children. By harnessing the power of PLM detection and innovative symbolic representation and phenotyping biomarkers, we have unveiled new dimensions in the objective and quantifiable assessment of children's speech and achieved the AUC of 0.71 and recall of 71.5% in the E2E TBD SBS SVM model. Our dual-method approach for core PLM detection, encompassing both acoustic and transcriptional analyses, has enabled a fine-grained, phoneme-specific evaluation that aligns with naturalistic speech patterns, especially in the spontaneous speech scenario. The statistical significance of our phenotyping features (such as MPD, LCC, ACC, and LCE) demonstrated through extensive analysis confirms the robustness and reliability of our E2E framework. Furthermore, the fine-tuning of our CTC model on the L2-ARCTIC dataset and the validation through five-fold cross-validation on both PLM detection and E2E SLI screening underscore the accuracy and generalizability of our screening pipeline. As we reflect on our contributions, we stand on the cusp of a new era in pediatric speech therapy, one that is informed by precise, reliable, and technologically

enriched tools designed to better the lives of children with SLI. Our work not only offers a forward-thinking solution for SLI screening but also sets the stage for future innovations that can build upon our foundational research.

REFERENCES

[1] P. Simon and S. Rosenbaum, "Speech and language disorders in children: Implications for the social security administration's supplemental security income program," 2016.

[2] H. M. Sharp and K. Hillenbrand, "Speech and language development and disorders in children," *Pediatric Clinics of North America*, vol. 55, no. 5, pp. 1159–1173, 2008.

[3] N. I. of Health, "Specific language impairment," https://www.nidcd.nih.gov/sites/default/files/Documents/health/voice/specific-language-impairment.pdf, July 2019.

[4] M. Shahin, U. Zafar, and B. Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2019.

[5] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigao, "Mispronunciation detection in children's reading of sentences," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1207–1219, 2018.

[6] M. Shahin, "Automatic screening of childhood speech sound disorders and detection of associated pronunciation errors," Ph.D. dissertation, UNSW Sydney, 2023.

[7] Y. Sharma and B. K. Singh, "One-dimensional convolutional neural network and hybrid deep-learning paradigm for classification of specific language impaired children using their speech," *Computer Methods and Programs in Biomedicine*, vol. 213, p. 106487, 2022.

[8] A. Hair, G. Zhao, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions." in *Interspeech*, 2021, pp. 2936–2940.

[9] P. A. Prelock, T. Hutchins, and F. P. Glascoe, "Speech-language impairment: how to identify the most common and least diagnosed disability of childhood," *The Medscape Journal of Medicine*, vol. 10, no. 6, p. 136, 2008.

[10] A. S.-L.-H. Association *et al.*, "Definitions of communication disorders and variations," 1993.

[11] R. E. Stoeckel, R. C. Colligan, W. J. Barbaresi, A. L. Weaver, J. M. Killian, and S. K. Katusic, "Early speech-language impairment and risk for written language disorder: A population-based study," *Journal of developmental and behavioral pediatrics: JDBP*, vol. 34, no. 1, p. 38, 2013.

[12] T. D. of Education, "Speech or language impairment evaluation guidance," 2009.

[13] A. S.-L.-H. Association, "Phonological processing," https://www.asha.org/practice-portal/clinical-topics/written-language-disorders/phonological-processing.

[14] R. Paul, *Language disorders from infancy through adolescence: Assessment & intervention*. Elsevier Health Sciences, 2007, vol. 324.

[15] V. Muter, C. Hulme, M. J. Snowling, and J. Stevenson, "Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study." *Developmental psychology*, vol. 40, no. 5, p. 665, 2004.

[16] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.

[17] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[18] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," 2009.

[19] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[20] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[21] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[22] L. Peng, Y. Gao, R. Bao, Y. Li, and J. Zhang, "End-to-end mispronunciation detection and diagnosis using transfer learning," *Applied Sciences*, vol. 13, no. 11, p. 6793, 2023.

[23] M. Black, J. Tepperman, S. Lee, P. Price, and S. S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[24] J. Duchateau, L. Cleuren, P. Ghesquière *et al.*, "Automatic assessment of children's reading level," in *Proceedings of the European Conference on Speech Communication and Technology*, 2007, pp. 1210–1213.

[25] E. Yilmaz, J. Pelemans, and H. Van Hamme, "Automatic assessment of children's reading with the flavor decoding using a phone confusion model," 2014.

[26] J. Paradis, P. Schneider, and T. S. Duncan, "Discriminating children with language impairment among english-language learners from diverse first-language backgrounds," 2013.

[27] L.-Y. Guo and P. Schneider, "Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 2, pp. 317–329, 2016.

[28] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," *arXiv preprint arXiv:2104.04045*, 2021.

[29] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[30] C. M. S. Group, "The cmu pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[33] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke *et al.*, "librosa/librosa: 0.10.1," https://doi.org/10.5281/zenodo.8252662, Aug. 2023.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[35] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[36] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[37] Rhasspy, "Github - rhasspy/gruut-ipa: Python library for manipulating pronunciations using the international phonetic alphabet (ipa)," https://github.com/rhasspy/gruut-ipa, Nov 2023.

[38] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[39] D. E. Knuth, "Two notes on notation," *The American Mathematical Monthly*, vol. 99, no. 5, pp. 403–422, 1992.

[40] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus." in *Interspeech*, 2018, pp. 2783–2787.

[41] V. Phy, "Automatic phoneme recognition on timit dataset with wav2vec 2.0." https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme, May 2023.