Automatic Speech Disorder Assessment for Children's Speech Disorder

by

Yaoxuan Luan

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 10, 2025

Keywords: Speech Processing, Neural Network, SSL Models,Embedding Distance, Speech Sound Disorder, AI

Approved by

Cheryl Seals, Chair, Professor of Computer Science and Software Engineering
Sathyanarayanan Aakur, Assistant Professor of Computer Science and Software Engineering
Gerry Dozier, Professor of Computer Science and Software Engineering
Marisha Speights, Assistant Professor in the Communication Sciences & Disorders Department, Northwestern University
Yang Zhou, Associate Professor of Computer Science and Software Engineering

Abstract

Speech disorders in children present persistent challenges for early detection and intervention due to the complex, variable, and context-dependent nature of developing speech. Traditional automatic speech disorder detection (ASDD) systems, which rely heavily on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), often struggle to capture the nuanced articulatory and prosodic patterns that characterize pediatric speech impairments. Recent advances in transformer-based deep learning architectures and self-supervised learning (SSL) offer promising alternatives for building more robust and interpretable ASDD systems.

This dissertation investigates three complementary approaches to advancing ASDD through the integration of modern representation learning techniques. The first study examines the use of the Vision Transformer (ViT) architecture applied to MFCC features for the classification of disordered and non-disordered child speech. By leveraging the ViT's patch-based attention mechanism, the study demonstrates that transformer-based models can achieve improved performance over conventional machine learning classifiers when applied to fixed acoustic feature representations.

The second study evaluates the effectiveness of SSL-based speech representations, specifically those derived from wav2vec 2.0 and HuBERT, in detecting speech disorders in children. Through layer-wise analysis and speaker-independent classification experiments, this study confirms that SSL representations outperform MFCCs by capturing more detailed, context-aware acoustic cues.

The third study explores an SSL-based perceptual similarity framework for measuring acoustic distances between speech samples. Using dynamic time warping (DTW) in the high-dimensional embedding space produced by SSL models, the study calculates similarity

scores between utterances without relying on textual transcriptions. These distance metrics are shown to correlate strongly with clinical judgments of speech pronunciation accuracy and disorder severity, supporting their potential use in continuous monitoring or pre-diagnostic screening.

Together, these studies provide a comprehensive evaluation of transformer-based and SSL-driven approaches for pediatric ASDD. The results highlight the advantages of using deep contextualized speech representations in terms of classification accuracy, robustness, and interpretability. The contributions offer a foundation for developing clinically viable tools to support early identification and longitudinal assessment of speech sound disorders in children.

Acknowledgments

This dissertation represents not only the culmination of my academic journey at Auburn University but also the collective effort of many individuals who supported me along the way. I am profoundly thankful to my advisor for her mentorship and guidance, my collaborators for their insight and encouragement, and my family for their unwavering love and belief in me. Their support made this work possible.

I want to express my deepest gratitude to my advisor, Dr. Cheryl Seals, whose guidance, insight, and steadfast support have been instrumental throughout my doctoral research and since the beginning of my graduate journey. Her unwavering encouragement, high standards, and belief in my potential pushed me to grow academically and professionally. She challenged me to think critically, supported me through every obstacle, and provided a model of leadership and integrity that I will carry forward in my career. I am truly fortunate to have had her mentorship across these formative years.

I am also sincerely thankful to my co-chair, Dr. Marisha Speights, for her thoughtful advice and encouragement, as well as to my committee members, Dr. Gerry Dozier, Dr. Sathyanarayanan Aakur, and Dr. Yang Zhou, for their invaluable feedback, constructive critiques, and generous investment of time and expertise. Each of them has played a vital role in helping me refine and strengthen this dissertation. I am especially grateful to Dr. Aurora Weaver for serving as my university reader. Her careful review and thoughtful comments have been incredibly helpful in shaping the final version of this work.

My deepest thanks go to my parents, Shaocheng Luan and Guoling Yang, whose unwavering love, sacrifices, and belief in me have been my greatest source of strength. They nurtured my curiosity and supported my dreams from the very beginning, even when the path was uncertain. Their support has been invaluable in my academic journey.

To my wife, Yanqi Zhang, I cannot fully express my gratitude for your endless patience, encouragement, and love. Your steadfast presence through every high and low has sustained me in ways I will always cherish. This achievement is not mine alone; it belongs to you, too.

War Eagle!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Speech Sound Disorders (SSDs) represent some of the most prevalent communication impairments among children, impacting approximately 8–9% of early school-age populations. If left unaddressed, these disorders may result in enduring repercussions on academic achievement, social integration, and long-term cognitive development. Therefore, the early detection and continuous monitoring of speech disorders are crucial for timely and effective intervention. However, the assessment of pediatric speech remains a highly specialized and resource-intensive task, usually requiring the expertise of trained clinicians and extensive manual transcription or perceptual scoring procedures.

Automatic speech disorder detection (ASDD) offers a scalable solution to reduce clinical workload and improve access to early screening. Conventional ASDD systems have largely relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) or prosodic cues to model speech deviations. Although effective in certain limited contexts, these features frequently overlook the complete spectrum of articulatory and phonetic variability in children's speech, which is naturally more inconsistent and less organized than that of adults. This shortcoming is especially apparent when analyzing disordered speech, which diverges markedly from standard acoustic and phonological patterns. As a result, conventional ASDD systems often experience challenges with generalizability and diminished interpretability in clinical settings.

Recent developments in deep learning, especially in transformer architectures and self-supervised learning (SSL), have brought about major advancements in speech representation. Models like wav2vec 2.0 and HuBERT utilize extensive, unlabeled speech datasets to acquire

rich, contextual embeddings directly from raw audio waveforms. These models have reached state-of-the-art performance across a variety of speech processing tasks and demonstrate potential in capturing subtle phonetic and prosodic features, presenting a strong alternative to manually crafted features for ASDD. Nevertheless, the use of SSL models for disordered child speech and their interpretability within clinical settings is still inadequately investigated.

1.2 Research Objectives

The primary aim of this dissertation is to enhance the detection of speech disorders in children using deep transformer-based techniques and self-supervised representation learning methods. This aim is pursued through these key objectives:

- To evaluate the performance of transformer-based classifiers, such as the Vision Transformer (ViT), when applied to handcrafted acoustic features for ASDD.
- To investigate the phonetic relevance, classification accuracy, and interpretability of SSL-based speech representations (wav2vec 2.0 and HuBERT) in detecting disordered speech.
- To propose and validate a perceptual similarity framework that estimates speech pronunciation and disorder severity using SSL-based acoustic distance metrics, without requiring transcription.

1.3 Dissertation Overview and Contributions

This dissertation is organized into three core studies, each addressing a distinct yet complementary aspect of speech disorder modeling:

- **Study 1** explores using the Vision Transformer (ViT) architecture for ASDD based on MFCC features. By treating MFCC spectrograms as image-like inputs, the study leverages ViT's attention mechanisms to model local and global acoustic dependencies, achieving improved classification performance over conventional baselines.

- **Study 2** conducts a layer-wise and speaker-independent evaluation of SSL models (wav2vec 2.0 and HuBERT) for ASDD. This study demonstrates the superior performance of SSL embeddings in capturing phonetic and prosodic variation and introduces a novel interpretability approach that links model saliency with phoneme-level forced alignment.

- **Study 3** proposes a transcription-free perceptual similarity framework using dynamic time warping (DTW) on SSL-derived embeddings (Wav2vec2) to measure speech distance. The resulting metrics show strong correlations with phonetic distances and speech pronunciation scores, providing a scalable automatic speech pronunciation assessment method for children's disordered speech.

The key contributions of this dissertation include:

- A systematic comparison between handcrafted and self-supervised speech features for ASDD.

- The adaptation and application of transformer models to pediatric speech data.

- An interpretable ASDD framework that integrates phoneme-level analysis with model saliency.

- A novel embedding-based distance metric for perceptual similarity and speech pronunciation scoring.

1.4 Significance and Impact

By bridging cutting-edge machine learning methods with clinically relevant tasks, this work contributes both theoretically and practically to the field of speech and language pathology. It offers novel computational tools for early identification and longitudinal monitoring of SSDs in children, potentially improving access to speech care in under-resourced settings.

Furthermore, the insights gained from this research provide a deeper understanding of how deep neural representations encode speech variability—laying the groundwork for more interpretable and personalized ASDD systems in the future.

Chapter 2

Study 1 - Detection of Children's Speech Disorder through Neural Network

2.1 Introduction

Speech disorders are categorized as communication impairments that affect a person's ability to produce speech sounds correctly and fluently. These may include difficulties in articulation, voice quality, resonance, or speech fluency. In the United States alone, it is estimated that nearly 2% of children between the ages of 3 and 17 are affected by some form of speech disorder. Among pediatric populations, the variability in speech patterns, pronunciation development, and co-occurring cognitive or developmental conditions makes disorder detection particularly challenging. Unlike adult speech, which is relatively stable and well-formed, children's speech often exhibits greater phonetic diversity and inconsistency. This variability increases the complexity of creating reliable automatic detection systems. Early detection is critical, as untreated speech disorders can significantly impact academic development, social interaction, and emotional well-being. However, many existing speech analysis systems are not yet accurate enough for clinical use. Traditional machine learning approaches, relying on handcrafted features and shallow classifiers, often struggle to exceed 75% accuracy, which is insufficient for reliable deployment in real-world pediatric settings.

To address this important clinical and technical challenge, I developed and evaluated a Transformer-based model for binary speech classification, aiming to distinguish between disordered and non-disordered children's speech. Specifically, I employed a Vision Transformer (ViT), a variant of the transformer architecture originally developed for computer vision tasks, to process MFCC representations of audio data. Transformers are known for their ability to capture long-range dependencies and contextual relationships within data,

which makes them particularly well-suited for analyzing the temporal and spectral structure of speech signals. Prior studies have shown that transformer models, when pre-trained on large-scale image data and adapted for audio, can perform effectively in domains such as speech emotion recognition and paralinguistic analysis. Leveraging this foundation, I adapted the ViT architecture to process 2D time–frequency representations of children's speech, treating MFCC matrices as image-like inputs.

This experimental pipeline included the use of two publicly available continuous speech datasets—Beginner Intelligibility Test (BIT) and Brown Bear, Brown Bear (BB)—which include speech samples from children with and without diagnosed speech disorders. I applied transfer learning by initializing our ViT model with pre-trained weights and fine-tuning it on our domain-specific classification task. The results of our preliminary study show that the ViT model achieved evaluation accuracies exceeding 91% on both datasets, along with strong specificity and sensitivity, outperforming previously reported benchmarks in the domain. These findings validate the hypothesis that transformer-based architectures can effectively generalize across diverse pediatric speech patterns and reliably distinguish disordered speech from typical speech with minimal supervision. Furthermore, the high F1 scores indicate that the model performs well across both precision and recall dimensions, which is essential in clinical applications where both false positives and false negatives carry significant consequences.

The broader significance of this work lies in its potential to bridge the gap between machine learning research and real-world clinical practice. By delivering a highly accurate and generalizable model, this research provides a step toward building an automated speech assessment tool that could assist speech-language pathologists (SLPs) in early diagnosis, triaging, and longitudinal monitoring of children with speech disorders. The transformer-based model not only offers a powerful alternative to conventional CNNs and RNNs, but also supports the development of more scalable, adaptable, and interpretable solutions for speech pathology. In conclusion, this study contributes to advancing the field of pediatric

speech technology and underscores the promise of deep learning models—particularly transformers—for improving clinical assessment workflows and enabling earlier, more objective diagnostic support for children with speech impairments.

2.2 Related works

2.2.1 Literature Reviews

Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders [24]

Speech is a fundamental behavior of humans, and voice becomes its primary subsystem. A person's regular speaking voice is created when the larynx works together with the pulses of air from the lungs to make the vocal folds move toward the center. However, a speech disorder is defined as any kind of abnormality that deviates from acoustic characteristics, such as 'loudness, pitch, and / or vocal flexibility', from typical vocal patterns of individuals of the same age, gender, and social group. [10] In clinical assessments, the perceptions of hearing of speech language pathologists are usually the ultimate evaluation of many speech disorders, especially for some components of complex communication disorders. In this paper, the author clarified that the weaknesses of auditory-perceptual judgments of speech need to be made known.

In this article, the author described the problems of auditory-perceptual judgments in clinical practice as follows:

· "Judges do not appear to have equivalent definitions of dimensions to be rated.

· Specialists fail to determine which perceptual dimensions should be rated for a given disorder.

· Perceptual ratings of various dimensions are intercorrelated; they are not independent. When this happens, the values obtained for any one dimension may be influenced by concurring dimensions of a disorder.

· Various perceptual dimensions are not rated with uniform reliability.

· Differences among expert judges are larger than the differences needed for diagnostic classification or the effects of intervention. ’

The author demonstrated several related experiments on speech perception to show many limitations in the accuracy of perceptual processing. When judging disordered voices, for example, clinicians do not have clear enough voice rating standards to scale voice quality. Another obstacle in voice rating is that judges might disagree with each other while rating disordered voices. Auditory perceptual assessment is a critical component of clinical evaluations in speech pathology, as it provides essential evidence for clinical decision-making.

In this paper, the author clarified that being aware of auditory-perceptual judgment errors and biases during clinical speech evaluation is an essential step in the practical and refined use of perceptual methods.

Dysarthria Detection Using Convolution Neural Network [35]

The authors presented an innovative application of Convolutional Neural Networks (CNNs) in detecting dysarthria, a motor speech disorder characterized by difficulty controlling muscles used in speech production. The results of this study highlighted the importance of early and accurate detection of dysarthria, given its association with numerous underlying neurological conditions, including stroke, Parkinson’s disease, and amyotrophic lateral sclerosis (ALS). The primary objective was to develop an effective and automated system capable of distinguishing dysarthric speech from healthy speech samples using CNN-based methods.

To achieve their goal, the authors leverage a combination of acoustic feature extraction and CNN architectures. Specifically, they extract multiple acoustic features known for their effectiveness in speech analysis, such as Mel-Frequency Cepstral Coefficients (MFCC),

zero-crossing rates, spectral centroid, and spectral roll-off, demonstrating comprehensive consideration of important speech descriptors. These features serve as critical input to CNN, helping the model to learn distinctive patterns associated with dysarthric speech.

The dataset employed for training and validating the model is the TORGO database, a well-established resource that includes speech samples from individuals with dysarthria and healthy control subjects. Given the limited size and variety in the available data, the authors utilize robust data augmentation techniques. They introduce Gaussian noise and apply low-pass filtering with a 4000 Hz cutoff to generate additional training samples, thus enhancing the model's generalization capability and robustness against variations in real-world data conditions.

The architecture described in the paper was designed to optimize feature learning and classification. It comprises multiple convolutional layers with rectified linear unit (ReLU) activations, followed by max-pooling layers to effectively reduce dimensionality and maintain essential speech features. Furthermore, dropout layers are integrated to prevent overfitting, a crucial consideration given the small size of the dysarthric speech dataset. The final stages of the architecture include fully connected dense layers with sigmoid activation, functioning effectively as a binary classifier to differentiate between healthy and dysarthric speech. Specifically, the convolutional layers perform the essential role of extracting nuanced acoustic characteristics, whereas the fully connected layers execute the classification by analyzing learned features.

A noteworthy aspect of this approach is its training process, which emphasizes both convolutional feature extraction and the densely connected classification layers. The authors clearly detail their training methodology, using standard practices such as backpropagation, mean square loss calculation, and appropriate learning rate adjustments. Their approach reflects a systematic effort to leverage CNN capabilities to identify subtle patterns in speech signals associated with dysarthria.

The reported results were particularly impressive, with the model achieving an accuracy of 93.87% on the TORGO dataset. Such results not only demonstrate the effectiveness of CNN-based approaches but also significantly outperform previous research efforts in dysarthria detection cited within the literature review section. This performance indicated the potential practical applicability of CNNs in clinical speech diagnostics and emphasizes the effectiveness of CNN architectures combined with relevant acoustic features.

However, despite the strong results and thoughtful design, several aspects warrant further exploration and refinement. Firstly, the authors focused on a single dataset, TORGO, which, while valuable, may limit the generalizability of their model across different languages, speech contexts, and disorder severity levels. Testing and validating the model on more diverse datasets (such as UA-Speech, home-recorded speech samples, or multilingual speech datasets) could significantly strengthen the findings.

Secondly, although the authors described the general architecture and training procedures, further elaboration on hyperparameter tuning and the rationale for specific architectural choices (e.g., kernel sizes, pooling methods, and dropout rates) would improve the methodology's clarity and reproducibility. Extensive ablation studies or sensitivity analyses could also clarify the impact of individual CNN components on performance.

Lastly, although acoustic features like MFCC, spectral centroid, and spectral roll-off are robust, exploring advanced feature extraction methods, such as self-supervised learning (SSL)-based embeddings from wav2vec 2.0 or HuBERT, could potentially yield superior performance due to their richer and more generalized speech representations. Integration or comparative analyses of CNN methods with SSL embeddings could provide a deeper understanding of optimal speech representations for dysarthria detection.

In conclusion, this work made a significant contribution to automatic speech disorder detection, demonstrating the strong potential of CNN-based approaches. This paper provided a solid methodological foundation and compelling empirical results that support further investigation and refinement of CNN approaches for automated dysarthria detection.

An Enhanced Speech Emotion Recognition Using Vision Transformer[1]

The authors showed a compact yet powerful framework for speech emotion recognition (SER), leveraging the capabilities of Vision Transformers (ViT) to process emotional cues embedded in mel-spectrogram representations of speech. Traditionally, SER systems have relied heavily on handcrafted acoustic features such as pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs), which often struggle to generalize in noisy environments or capture the nuanced patterns of affective speech. To overcome these limitations, the authors converted raw audio signals into mel-spectrograms, two-dimensional (2D) time–frequency visual representations that reflect how energy is distributed over time and frequency. Then they fed these visual maps into a lightweight Vision Transformer (ViT) architecture. The core innovation lay in how the ViT processed the spectrogram: the model treated the mel-spectrogram as an image, divided it into fixed-size non-overlapping patches (in this case, 32×32), flattened each patch, and linearly projected them into a latent embedding space of 128 dimensions. These patch embeddings were then enriched with positional encoding, which allowed the model to maintain awareness of the spatial location of each patch within the overall spectrogram.

Unlike CNNs, which primarily learn local features through convolutional filters, ViT uses a self-attention mechanism to learn global relationships among all patches in the image. This is particularly beneficial for SER, as emotional cues in speech often span multiple time–frequency regions that are not spatially adjacent. The self-attention module calculates similarity scores between all patch pairs via dot-product attention, allowing the model to dynamically focus on the most emotionally salient regions regardless of their positions in the input. Multiple attention heads are used in parallel to capture diverse dependencies, and the outputs are combined through multi-head attention. The transformer encoder then passes the representations through feedforward layers with Gaussian Error Linear Unit (GELU) activations, which are known to outperform ReLU in various speech and NLP tasks. Dropout and layer normalization are employed for regularization and training stability.

The ViT model was evaluated on the TESS and EMO-DB datasets, which contained speech samples labeled across seven emotional categories. The proposed model achieved state-of-the-art results — 98% accuracy on TESS and 91% on EMO-DB — and maintained 93% accuracy when the two datasets were merged, indicating strong generalizability. These results were bolstered by comprehensive ablation studies, which showed that removing key components such as dropout or altering the patch size reduced performance, thereby confirming the architectural choices. The authors also conducted comparative evaluations against deep CNN models such as ResNet, DenseNet, and MobileNet, demonstrating that their ViT-based architecture achieved superior accuracy with fewer parameters (4.17M), making it suitable for real-time deployment. A notable contribution was the model's ability to outperform CNNs in detecting difficult emotions like "disgust" and "neutral," which are often confused due to subtle prosodic variations.

In summary, this work demonstrated how a well-designed Vision Transformer could extract emotionally relevant features from mel-spectrogram images by leveraging global attention mechanisms, positional context, and parameter-efficient design. This marked a significant step forward in SER research by bridging techniques from computer vision and speech processing. While the system used only mel-spectrograms as input, future enhancements could involve integrating other acoustic features such as MFCCs, chromagrams, or temporal dynamics from raw waveforms, potentially enabling the model to perform well under spontaneous or noisy speech conditions. Nonetheless, the ViT-based approach already set a new benchmark for lightweight, accurate, and scalable SER systems suitable for real-world human–computer interaction applications.

2.2.2 Research Problems / Hypothesis

Understanding and detecting speech disorders in children presents unique challenges in the field of speech processing and machine learning. Existing approaches often rely on handcrafted acoustic features (e.g., MFCCs, formants, pitch) combined with conventional

classifiers (SVMs, HMMs, or CNN). However, these methods tend to underperform when applied to disordered speech, particularly in pediatric populations, due to the high variability, limited datasets, and the subtlety of the acoustic cues. The increasing availability of deep learning methods, particularly architectures based on self-attention mechanisms such as Transformers, opens a new direction for addressing these limitations. While Transformer models have demonstrated success in natural language processing and, more recently, computer vision, their potential in modeling the temporal–spectral structure of disordered speech remains underexplored.

This research aimed to investigate whether a Transformer-based model, adapted for audio input via spectrogram transformation, can offer superior performance in detecting children's speech disorders. Specifically, it explores whether the architectural advantages of Vision Transformers (ViTs)—such as their ability to model global dependencies and reduce inductive biases—can be harnessed for analyzing the rich time–frequency structure inherent in pathological speech.

Research Questions: The following research questions guide this study:

- **Model Design**: What kind of machine learning architecture—among CNNs, or Transformers—is best suited for capturing the complex acoustic and temporal patterns associated with disordered children's speech?
- **Evaluation**: What metrics are most appropriate for assessing the performance of models in detecting children's speech disorders? How can these metrics reflect clinical relevance?
- **Performance Benchmarking**: What is the achievable classification performance (e.g., in terms of accuracy, precision, recall, F1-score) using a Transformer-based model on disordered speech datasets?

Proposed solutions: This study proposes a novel Transformer-based architecture for detecting children's speech disorders. The core idea is to transform raw audio signals into

time-frequency representations using both mel-spectrograms and MFCCs. These representations are divided into patches and processed through a Vision Transformer (ViT) that uses self-attention to learn contextual relationships across the spectrogram space. This allows the model to capture subtle acoustic variations often indicative of speech impairment.

The model will incorporate standard audio preprocessing techniques including silence trimming, pre-emphasis filtering, and normalization. To ensure the model is both effective and efficient, training will use classification metrics including accuracy, precision, recall, and F1-score. Hyperparameters such as patch size and dropout rate will be tuned, and regularization techniques like early stopping will be employed to prevent overfitting.

Research Hypotheses:

- H1: A Vision Transformer architecture, when combined with appropriate audio preprocessing (MFCCs), can be effectively adapted for detecting speech disorders in children.
- H2: The proposed Transformer-based model will outperform traditional deep learning models (e.g., CNNs) in classifying disordered vs. typical children's speech by better modeling global contextual patterns.

2.3 Methods

2.3.1 Data

In this proposed experiment, I used two datasets from the Speech Exemplar and Evaluation Database (SEED) [25] database, which contains about 16,000 speech samples recorded by participants aged from 2 to 85 years. The SEED project aims to supply researchers and phonetic transcription training instructors with high-quality speech sample recordings. This initiative enhances the resources available for studying and teaching the nuances of speech sounds and patterns. Two datasets from the SEED, the Beginner Intelligibility Test (BIT) [27] and Brown Bear, Brown Bear (BB) [26] were selected for analysis. These contain

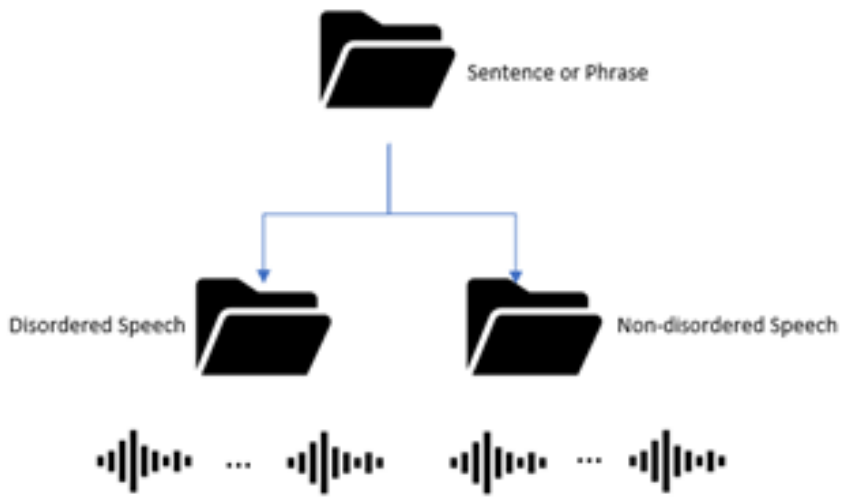

Figure 2.1: Enter Caption

Table 2.1: SEED Datasets

	Beginner Intelligibility Test (BIT)	Brown Bear, brown bear (BB)
Number of Sentences	40	33
Number of Audio	696	523

sentence-length recordings (n = 1,219) from 60 preschool children aged 3.0 to 5.0 years (n = 45 without speech disorders; n = 15 with speech disorders).

2.3.2 Mel-frequency Cepstral coefficients

The Mel Frequency Cepstral Coefficients (MFCCs) [22] are extensively utilized in human speech recognition due to their foundation in human auditory perceptions. MFCCs take into account the human auditory system's heightened sensitivity to frequencies between 1 kHz and 4 kHz, which are critical for speech perception. The **Figure 2.2** below shows the process of extracting MFCCs from a speech audio file. The process of extracting MFCCs from raw speech audio involves a sequence of signal processing steps designed to mimic the human auditory system's perception of sound. MFCCs are among the most commonly used features in speech processing due to their effectiveness in capturing the phonetic content of speech in a compact and discriminative form.

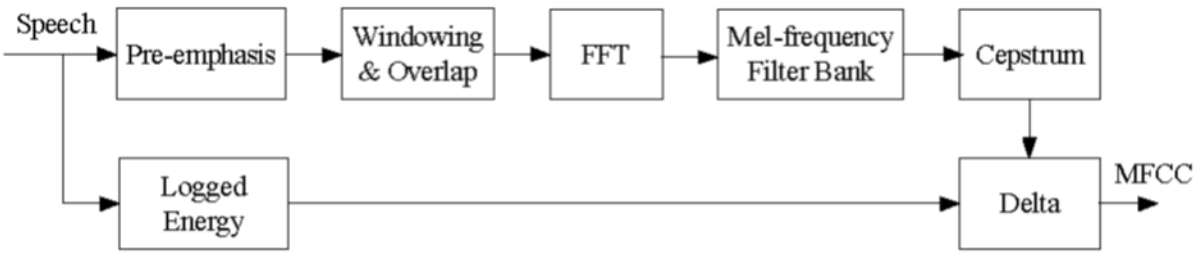

Figure 2.2: Process to extract MFCCs from raw speech audio file

1. **Raw Speech Input** The input is a time-domain waveform sampled at a fixed rate (e.g., 16 kHz or 44.1 kHz). Speech signals are non-stationary over long durations but can be considered stationary over short segments (20–40 ms), making frame-wise processing suitable.

2. **Pre-emphasis:** A first-order high-pass filter is applied to the speech signal to amplify high-frequency components that are often suppressed during sound production. This enhances the intelligibility and robustness of features. The filter is defined by:

$$y[n] = x[n] - \alpha x[n-1] \tag{2.1}$$

where α is typically 0.95 to 0.98. This step flattens the spectral envelope, equalizing the frequency content of the signal.

3. **Framing and Windowing** The pre-emphasized signal is divided into short overlapping frames (e.g., 25 ms duration with 10 ms overlap), allowing analysis of quasi-stationary speech segments. Each frame is multiplied by a **window function**, typically the **Hamming window**, to reduce spectral leakage and discontinuities at the frame edges:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{2.2}$$

This ensures smoother transitions and better spectral resolution.

4. **Fast Fourier Transform (FFT)** Each windowed frame is transformed from the time domain to the frequency domain using the **FFT**, resulting in a complex-valued frequency spectrum. The magnitude (or power) spectrum is retained to analyze the distribution of energy over frequency:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \tag{2.3}$$

   This provides the basis for further spectral analysis.

5. **Mel-Frequency Filter Bank** The magnitude spectrum is passed through a **mel-scale triangular filter bank**, where each filter is designed to approximate the critical bandwidths of human auditory perception. The mel scale is defined by:

$$m(f) = 2595 \cdot \log_{10}\left(1 + \frac{700}{f}\right) \tag{2.4}$$

   The filters are spaced linearly below 1 kHz and logarithmically above it. Each filter sums energy in its respective band, emphasizing perceptually important frequencies.

6. **Logarithmic Compression** To simulate the human ear's nonlinear sensitivity to loudness, the energy output from each mel filter is converted to a **logarithmic scale**:

$$E_{\log}(i) = \log(E(i)) \tag{2.5}$$

   This compression also improves numerical stability and emphasizes relative changes in energy, which are more informative than absolute magnitudes.

7. **Discrete Cosine Transform (DCT) — Cepstrum** The log-mel energies are decorrelated using the **DCT**, which compacts most of the signal energy into the first few coefficients. The resulting values are called **cepstral coefficients**, and typically the

first 12–13 coefficients are retained (excluding the 0th, which represents overall energy):

$$c[n] = \sum_{k=1}^{K} \log(E(k)) \cdot \cos\left[\frac{\pi n}{K}\,(k - 0.5)\right], \quad n = 1, 2, \ldots, N \tag{2.6}$$

8. **Delta and Delta-Delta Features** To capture the **temporal dynamics** of speech, first-order (**delta**) and second-order (**delta-delta**) derivatives of the cepstral coefficients are computed across frames:

$$\Delta c_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.7}$$

   These derivatives represent velocity and acceleration in cepstral space and are crucial for modeling transitions, coarticulation, and prosody. They are often appended to the static MFCCs, yielding a final 39-dimensional feature vector per frame (13 static + 13 delta + 13 delta-delta).

9. **Output: MFCC Feature Vectors** Each frame yields a feature vector, and the sequence of such vectors across the entire utterance constitutes the input to downstream classifiers (e.g., HMMs, SVMs, DNNs, Transformers). These MFCC features effectively represent the spectral and temporal structure of the speech signal in a compact and perceptually relevant way.

MFCC Visualization for Disordered vs. Non-disordered Speech

Figures 2.3 and 2.4 show the Mel-Frequency Cepstral Coefficient (MFCC) representations of two speech utterances—one from a **non-disordered (typical)** speaker and the other from a **disordered** speaker, respectively. These visualizations offer insights into the acoustic and articulatory characteristics of each class, serving as critical input features for automatic speech classification systems.

Figure 2.3: Non-disordered Speech

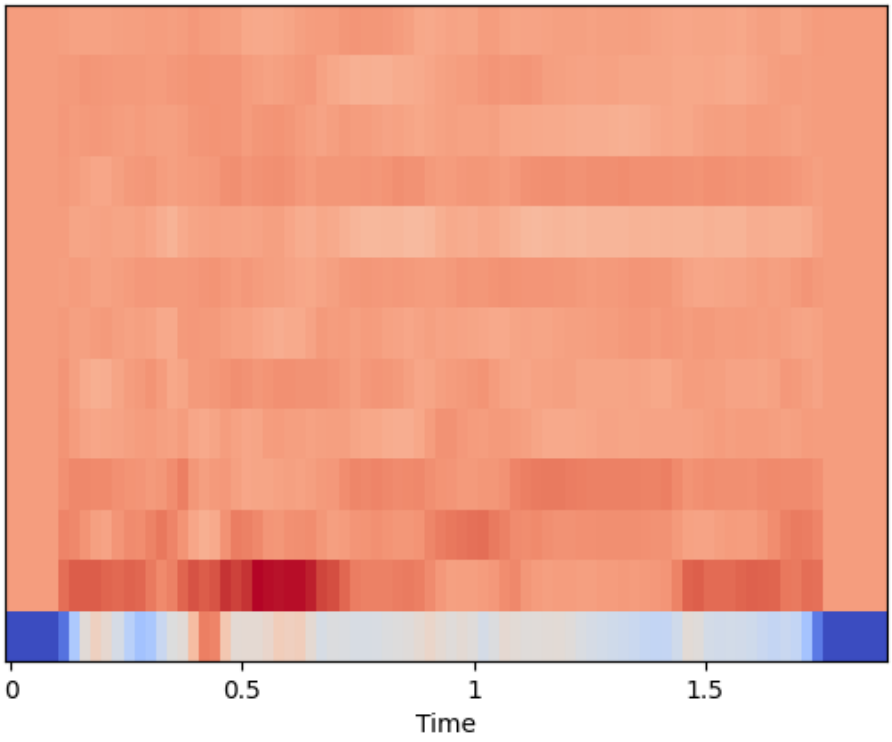

Figure 2.3: MFCC image of non-disordered speech

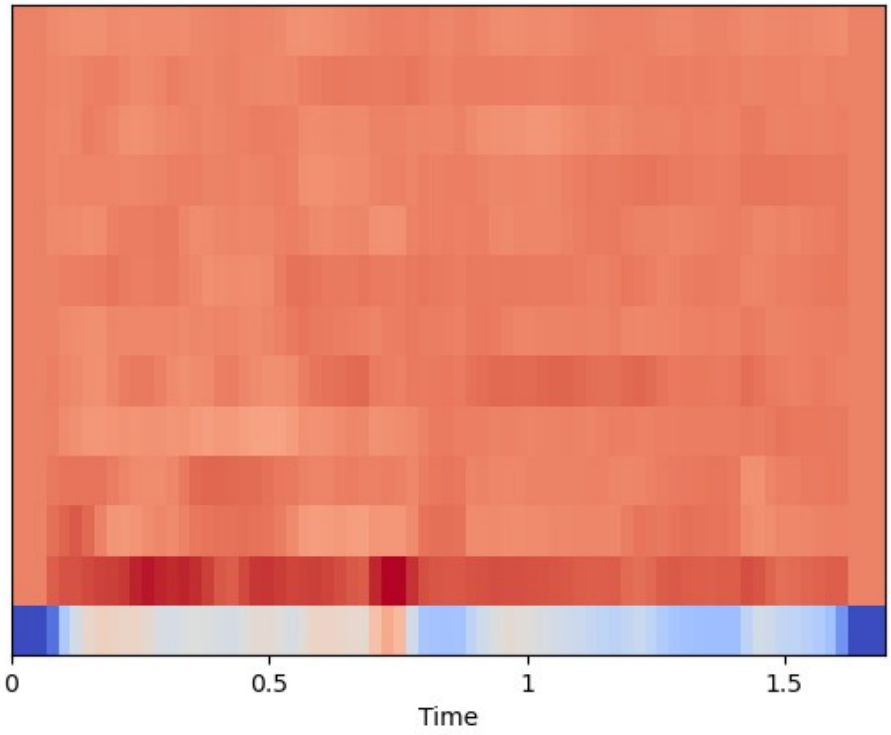

Figure 2.4: MFCC image of disordered speech

The MFCC map for typical speech reveals a **more stable and smoother temporal structure**, with clear vertical striations suggesting regular and predictable transitions between phonemes. This regularity reflects healthy speech production, characterized by consistent voicing, coarticulation, and formant trajectories. The color gradients are more uniform across time, indicating steady energy distribution over the cepstral bands.

Figure 2.4: Disordered Speech

In contrast, the disordered speech MFCC map shows **less consistent and more irregular patterns**. The time axis reveals frequent and abrupt shifts in energy distribution, especially in the lower cepstral bands, which may indicate instability in vowel production, slurring, or impaired motor control. There is a tendency for overlapping or smeared frequency content, which can result from atypical prosody, articulatory imprecision, or vocal tremor—common markers in speech pathology.

Clinical and Computational Implications

The visual and statistical differences in MFCC patterns between non-disordered and disordered speech are pivotal for downstream machine learning models, especially in classification tasks using SVMs, CNNs, or Transformers. By capturing temporal-spectral instability, MFCCs act as a discriminative feature space for voice pathology detection, as demonstrated in multitaper-based methods explored by [13].

2.3.3 Vision Transformer (ViT) Model

The Transformer model is a deep-learning neural network that was initially proposed by [49] in the Natural Language Processing (NLP) task. This innovative model stands out by not relying on traditional recurrent layers. The self-attention mechanism is pivotal to the functionality of the Transformer, enabling the decoder to process the entire input sequence simultaneously. This approach allows the model to focus on specific parts of the input that are most relevant for the task at hand, enhancing the efficiency and accuracy of decoding.

The architecture of the Transformer is characterized by a combination of self-attention and fully connected layers, incorporated into both the encoder and the decoder components. The encoder in the Transformer model comprises two key elements: a multi-head self-attention mechanism and a feed-forward neural network layer. The multi-head self-attention mechanism is designed to process the input from multiple perspectives simultaneously, allowing the model to capture a richer understanding of the input data. The feed-forward layer, on the other hand, processes the outputs of the self-attention mechanism to further refine the representation of the input sequence[49].

Similarly, the decoder in the Transformer model shares several structural similarities with the encoder. It includes two multi-head self-attention mechanisms and a single feed-forward layer. The first multi-head self-attention mechanism in the decoder is responsible for focusing on the appropriate parts of the input sequence, while the second is designed to ensure that the decoding process takes into account the entire context of the sequence. The feed-forward layer in the decoder then works to transform the attention-enhanced representations into the final output.

The scaled dot-product attention mechanism is a fundamental component in the architecture of the Transformer model, playing a critical role in its ability to efficiently handle sequence-to-sequence tasks. The inputs to the scaled dot-product attention are Queries (Q), Keys (K), and Values (V). These are typically matrices obtained by transforming the input data through learned weight matrices. Queries (Q) can be thought of as representations of the elements in the sequence for which I want to compute attention. In the context of a Transformer model, a query is associated with every element (e.g., a word in a sentence) that is seeking to find which parts of the input sequence are most relevant to it. Keys (K) are representations of the elements in the input sequence that are used to compute the attention mechanism. They pair with corresponding Values (V). Values (V) are also representations of the elements in the input sequence, but they are used in a different way compared to keys. Each value is associated with a key. Once the attention scores are computed (from

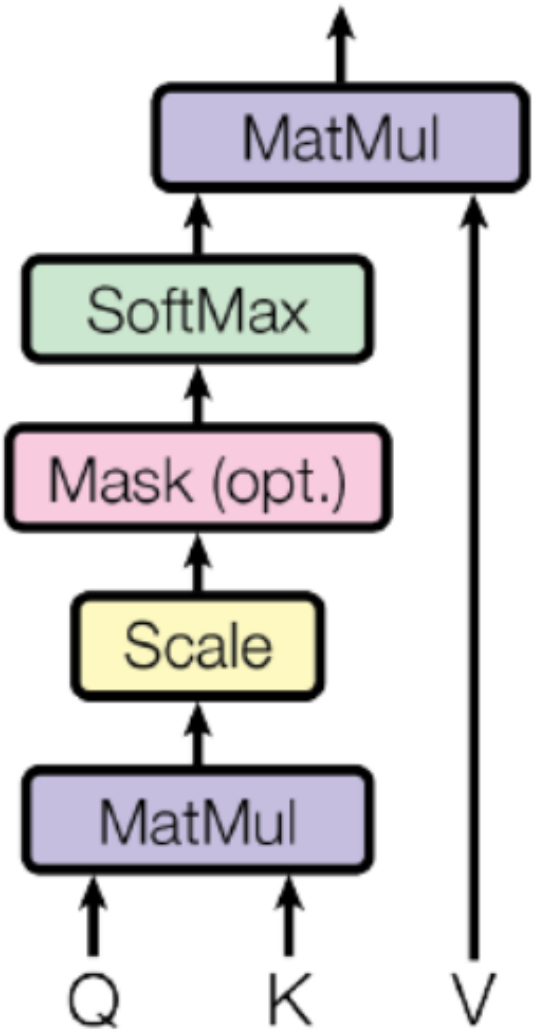

Figure 2.5: Scaled Dot-Product Attention

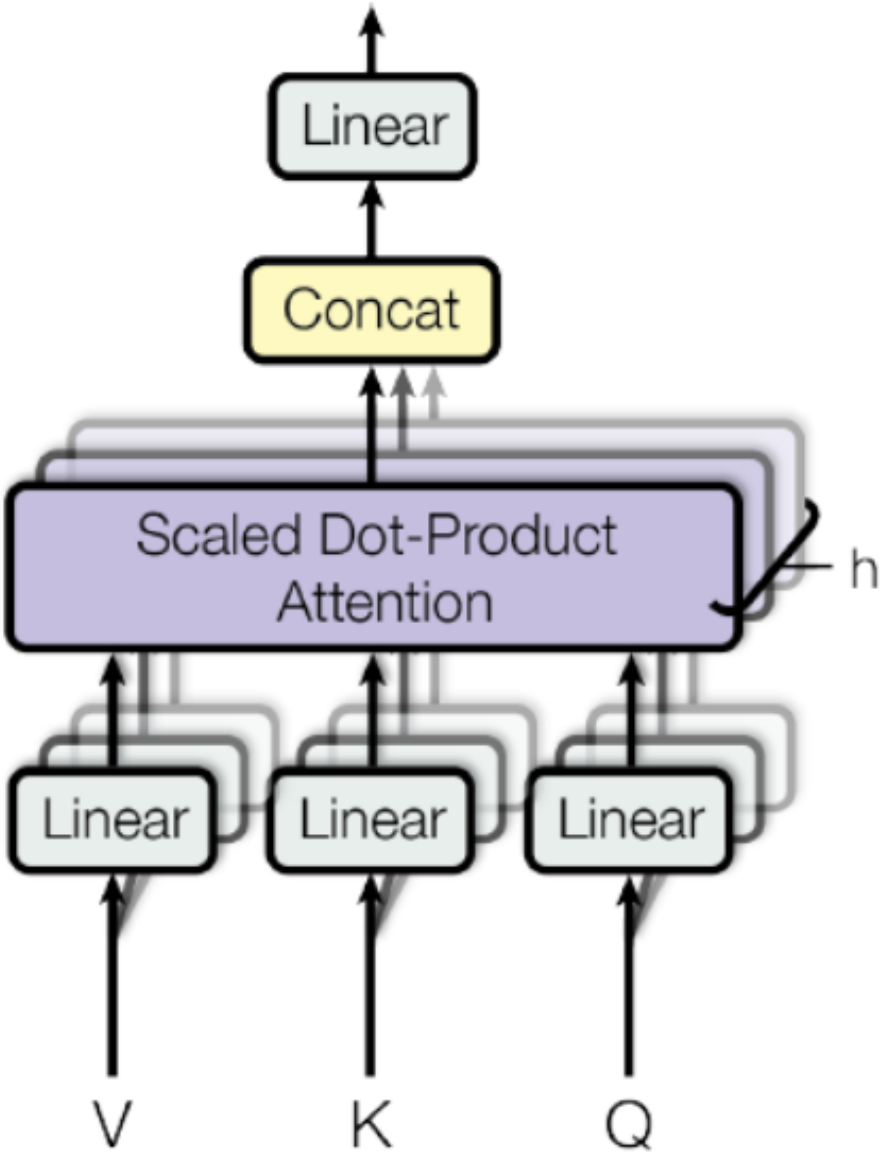

Figure 2.6: Multi-Head Attention consists of several attention layers running in parallel

Queries (Q) and Keys (K)), these scores are used to weigh the Values (V). The weighted sum of these Values (V), based on the attention scores, forms the output of the attention mechanism for each Queries (Q). The dot product scores are scaled down by dividing them by the square root of the dimensionality of the Keys (K). This scaling factor is crucial as it helps in stabilizing the gradients during training, particularly for higher dimensionalities where the dot products can be large, leading to a SoftMax function with extremely small gradients. So, I compute the matrix of output as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.8}$$

The scaled dot-product attention enables the model to dynamically weigh the significance of different parts of the input sequence. This adaptability is crucial for tasks that require an understanding of context, such as language translation or question answering. Unlike recurrent neural networks, the scaled dot-product attention allows for parallel processing of the sequence data. This parallelization is a significant advantage in terms of computational efficiency and training speed. Multi-head attention (**Figure 2.6**) is a sophisticated extension of the basic attention mechanism, particularly prominent in the Transformer model, a state-of-the-art architecture used in many natural language processing tasks. In a multi-head attention mechanism, the attention process (using Queries (Q), Keys (K), and Values (V)) is replicated multiple times. Each replication is known as a 'head'. Each head has its own set of linearly transformed queries, keys, and values, obtained by applying different learned linear projections to the input. After each head computes its attention output, these outputs are concatenated and once again linearly transformed. This final transformation is a way to combine the different learned aspects from each head into a unified output. The multi-head attention function is following:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.9}$$

Where

$$\text{head}_h = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.10}$$

$QW_i^Q, KW_i^Q, VW_i^Q$are parameter matrices calculated from Queries (Q), Keys (K), and Values (V). In essence, the encoder's role is to map the input sequences into abstract, attention-based representations. These representations capture the nuances and dependencies within the input data. The decoder, utilizing these representations, generates the output by focusing on relevant parts of these attention-based representations. This process allows the Transformer to handle a wide range of sequence-based tasks with remarkable efficiency and accuracy, making it a groundbreaking model in the field of natural language processing and beyond. In this study, I proposed to build a Transformer-based Audio model to detect Children's speech disorders. Compared to the convolutional neural networks (CNNs), which have been widely used to learn representations from audio spectrograms [29]and also in speech disorder detection task [31]. Inspired by the achievements of models solely based on attention mechanisms in the field of computer vision [12], it is time to introduce the attention mechanism to the field of children's speech disorder detection.

To make the standard Transformer handle the multi-dimensional images, a previous study [12] proposed a method to reshape the image into a sequence of flattened 2D patches. In this study, I propose to split the MFCCs image into a sequence of patches, where N is the number of patches the effective input sequence length for the Transformer model. As I need the model to detect speech disorder or non-disorder, which is a binary classification, I only use the encoder of the standard Transformer model. I just use the encoder from the standard Transformer [49]. In particular, the Transformer encoder has an embedding dimension of 768, 12 layers, and 12 heads. The Transformer encoder outputs the [CLS] token, which serves as the audio MFCCs representation. A linear layer, coupled with a sigmoid activation function, transforms the representation of the audio MFCCs into labels suitable for classification purposes.

2.4 Implementation

In this proposed experiment, I used two datasets form the **SEED** database[45], which contains about 1,600 speech samples recorded by participants aged from 2 to 85 years. The two datasets, Beginner Intelligibility Test (BIT) and Brown Bear, Brown Bear (BB), were used for training and validation and together contain 1,049 speech audio files. To avoid overfitting, 80% of the dataset was used as the training set, and the remaining data was used for testing/validation. Following the approach described in [18], a Vision Transformer (ViT) model pre-trained on a large-scale image dataset was utilized [11].

All audio files were labeled and placed in separate directories. Using the Librosa library [36], each audio file was converted to MFCC features.

TensorFlow Keras[9] framework was utilized to construct the CNN architecture. I employed the Adam algorithm [25] for gradient-based optimization and utilized the binary cross-entropy loss function during training.

Convolutional Neural Network (CNN) is a deep learning model widely used for image classification and computer vision tasks. The CNN models are mainly composed of three types of layers [31]. As a comparison, here is the CNN model summary: The convolutional neural network (CNN) architecture shown in the figure was designed to operate on two-dimensional audio feature representations, such as Mel-Frequency Cepstral Coefficients (MFCCs) or mel-spectrograms, which are widely used in speech analysis tasks due to their ability to capture perceptually relevant time–frequency information. The model began with a two-dimensional convolutional layer (`conv2d_4`), which applied 32 filters of size 3×3 to the input spectrogram. This layer was responsible for extracting basic local patterns in the input, such as vertical or horizontal edges that represent sudden energy shifts across frequency or time. The output was then passed through a max pooling layer (`max_pooling2d_4`), which reduced the spatial resolution by a factor of two, enabling the network to generalize better by focusing on the most salient features while discarding noise and redundancy.

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 148, 148, 32)	896
max_pooling2d_4 (MaxPooling2D)	(None, 74, 74, 32)	0
conv2d_5 (Conv2D)	(None, 72, 72, 64)	18496
max_pooling2d_5 (MaxPooling2D)	(None, 36, 36, 64)	0
conv2d_6 (Conv2D)	(None, 34, 34, 128)	73856
max_pooling2d_6 (MaxPooling2D)	(None, 17, 17, 128)	0
conv2d_7 (Conv2D)	(None, 15, 15, 128)	147584
max_pooling2d_7 (MaxPooling2D)	(None, 7, 7, 128)	0
flatten_1 (Flatten)	(None, 6272)	0
dense_2 (Dense)	(None, 512)	3211776
dense_3 (Dense)	(None, 1)	513

Figure 2.7: CNN

Subsequent convolutional layers increased the number of filters to 64 and then to 128 in deeper layers, progressively enabling the model to learn more complex and abstract features, such as spectral shapes, phoneme transitions, and even speaker-specific or disorder-specific articulatory patterns. Each convolutional layer was followed by a max pooling operation, which halved the spatial dimensions (e.g., from 72×72 to 36×36), thereby reducing the number of parameters and computational complexity while preserving the most dominant activations. By the time the signal reaches the fourth convolutional block (`conv2d_7` and `max_pooling2d_7`), the feature map had been reduced to a compact representation of shape 7×7×128, where each of the 128 channels encoded high-level information extracted from the original speech input.

This compact yet information-rich tensor was then passed through a \verb—Flatten— layer, transforming it into a one-dimensional vector of 6,272 elements, which was required for connection to the dense (fully connected) layers. The subsequent dense layer (`dense_2`) contains 512 neurons and more than 3.2 million trainable parameters, making it the most computationally intensive part of the network. This layer integrated information from the entire spectrogram and acted as a high-capacity classifier that combined multiple features to form a robust decision function. It enabled the model to learn nuanced combinations of features that were often indicative of specific speech disorders, such as inconsistent vowel articulation, timing irregularities, or phoneme substitutions. The final output layer (`dense_3`) consisted of a single neuron, which output a scalar value representing the probability of a binary class (e.g., disordered vs. non-disordered speech) using a sigmoid activation function.

Figure 2.8 presents the architecture of a Transformer-based model specifically adapted for speech classification using Mel-Frequency Cepstral Coefficients (MFCCs) as input. The pipeline begins with a raw audio waveform, which undergoes preprocessing to extract MFCC features—a compact and perceptually motivated representation of the speech signal that preserves the spectral envelope while discarding pitch and other less critical variations. This MFCC matrix, essentially a 2D image where rows correspond to cepstral coefficients and

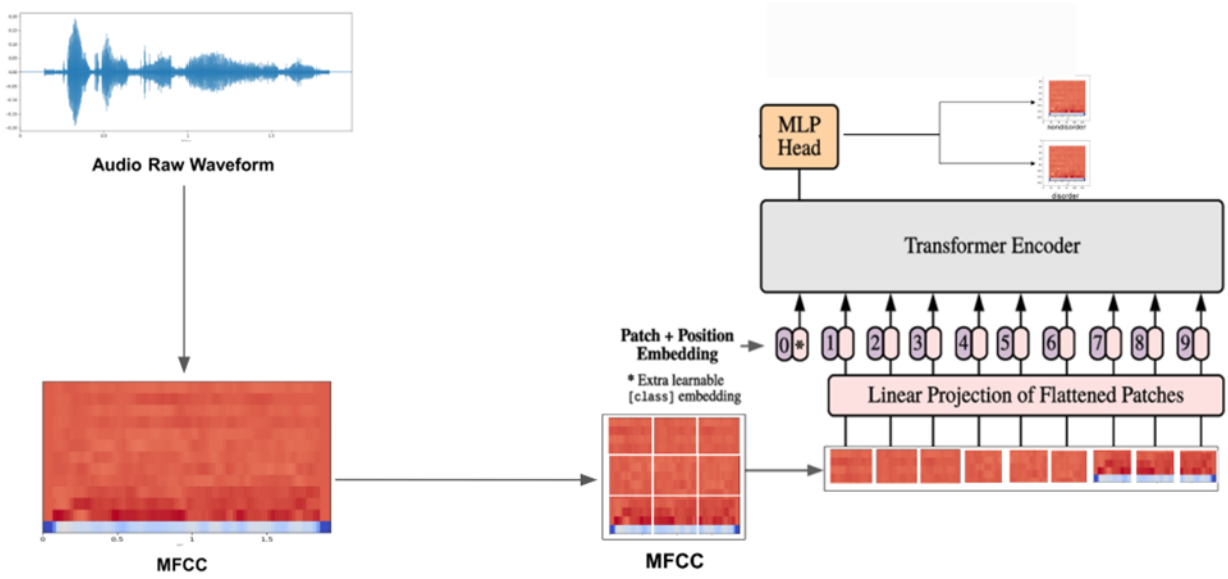

Figure 2.8: The Transformer-based speech model

columns represent successive time frames, is then segmented into fixed-size non-overlapping patches (e.g., 16×16). Each patch is flattened into a one-dimensional vector and mapped to a higher-dimensional embedding space via a learnable linear projection. This transformation prepares the input for processing by the Transformer, which requires a sequence of vector tokens rather than raw matrix input. To retain the positional context lost during flattening, a trainable **position embedding** is added to each patch embedding, encoding the temporal and spectral location of each patch. A special learnable `[CLS]` (classification) token is prepended to the sequence; this token is designed to aggregate global information across all patches during attention operations.

The sequence of patch embeddings, including the `[CLS]` token, was then passed into a **Transformer encoder**, composed of multiple layers of **multi-head self-attention**, **layer normalization**, and **feedforward neural networks**. The self-attention mechanism allows the model to compute relationships between all patches simultaneously, making it highly effective at capturing long-range dependencies in the time–frequency domain—an area where CNNs are typically limited by their local receptive fields. For instance, pathological speech characteristics such as inconsistent voicing, abnormal transitions, or prolonged phonemes

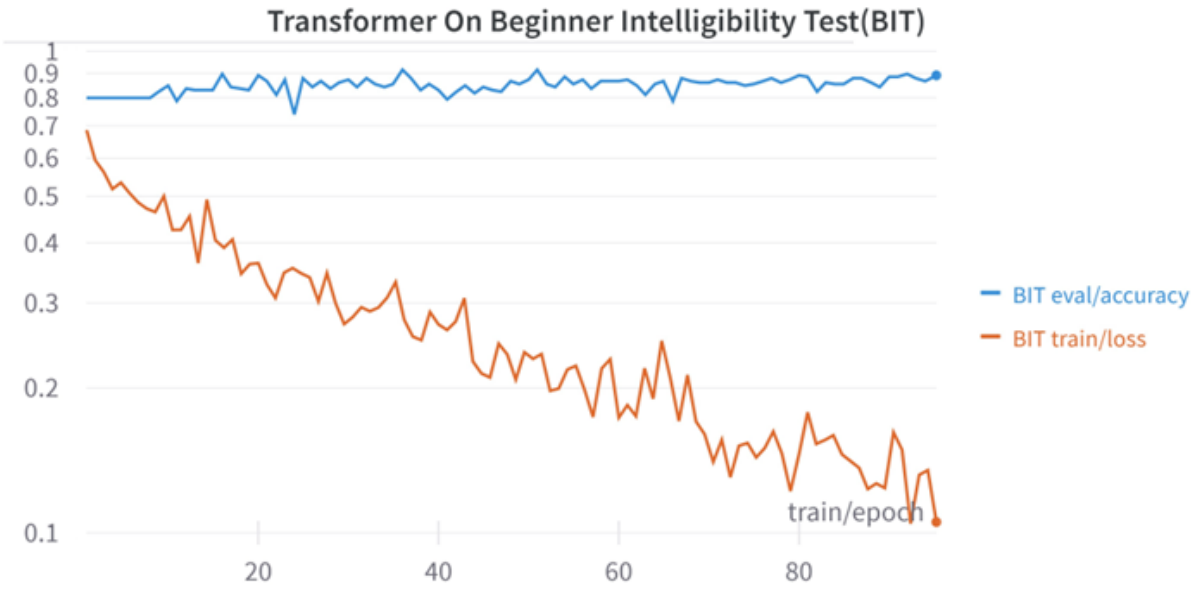

Figure 2.9: ViT on BIT

may span large temporal contexts, which the Transformer is well-suited to model holistically. The contextualized representation of the `[CLS]` token at the output of the encoder is passed through a **multi-layer perceptron (MLP) head**, which maps the learned global feature into a classification score. This score is then used to determine the final prediction, such as whether the input speech corresponds to a disordered or non-disordered class.

The Transformer library from Huggingface [52] is utilized to construct the ViT architecture. The Adam algorithm with a learning rate of $5e-05$ and cross-entropy was used during training.

We conducted our training on the Google Colab platform with T4 GPU. The previous study proves the Transformer can perform well on audio tasks with pre-trained image data [19], which can be applied to improve classification accuracy in children's speech disorders.

2.5 Results

The training curves shown in Figures 1.9 through 1.12 provide a comparative analysis of the learning behavior of Vision Transformer (ViT) and Convolutional Neural Network (CNN)

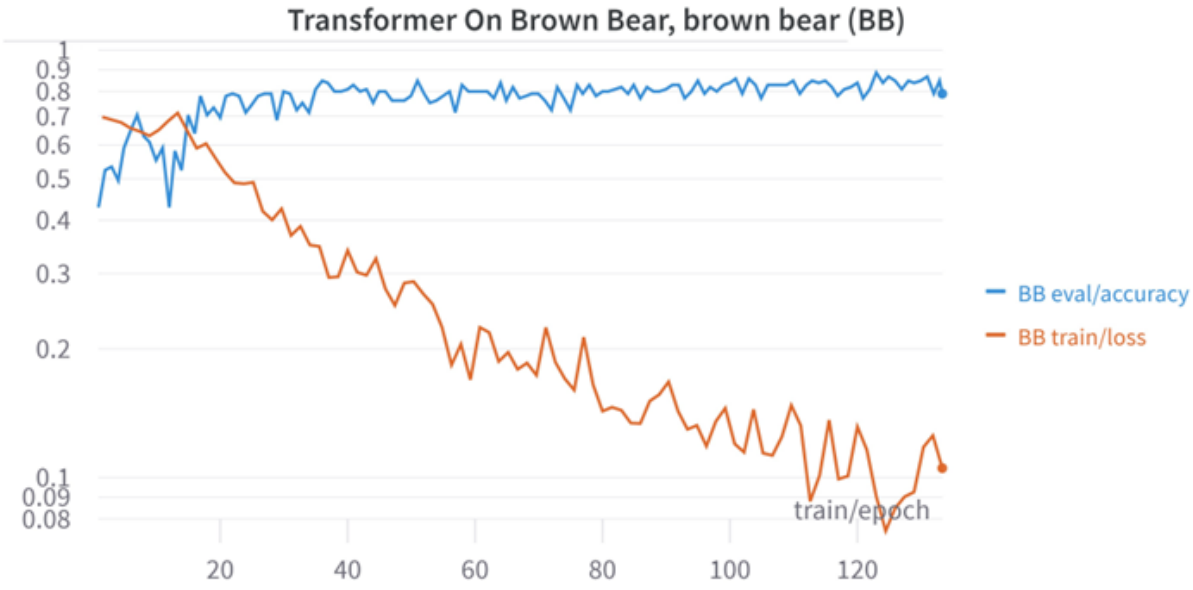

Figure 2.10: ViT on BB

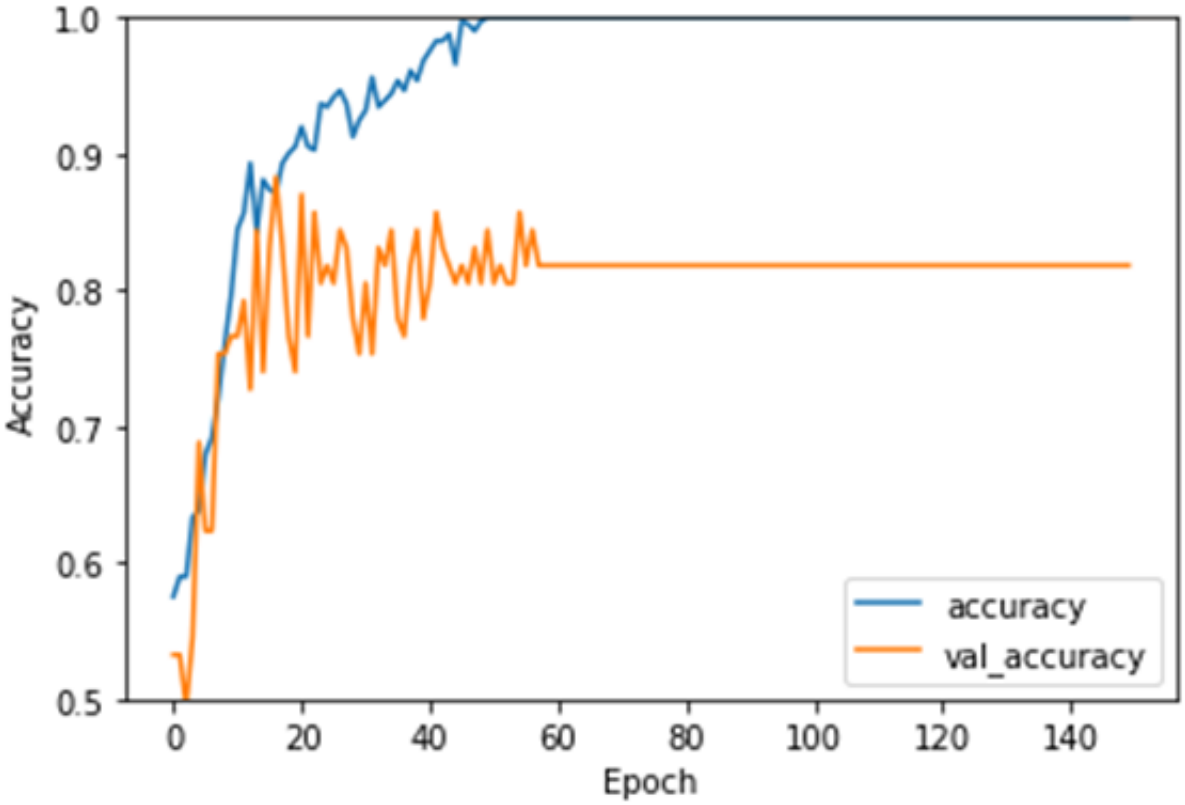

Figure 2.11: CNN on BIT

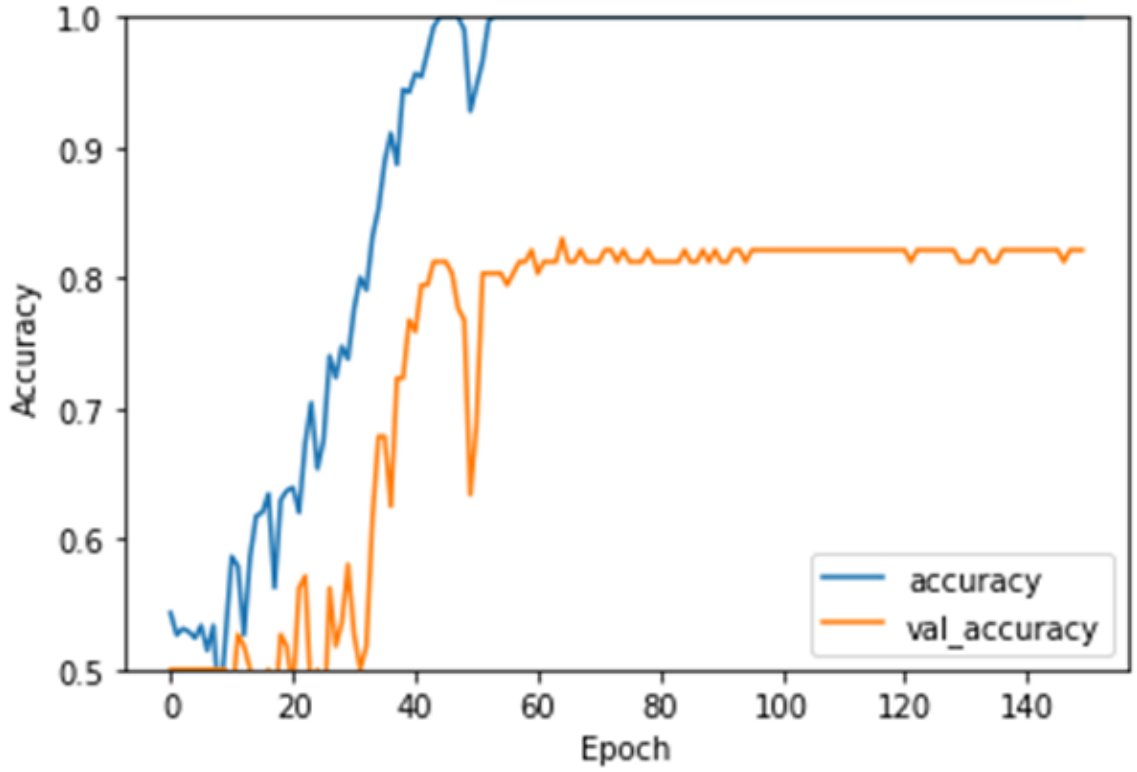

Figure 2.12: CNN on BB

models on two speech classification datasets: the Beginner Intelligibility Test (BIT) and the Brown Bear, Brown Bear (BB) dataset. In Figures 2.9 and 2.10, the ViT model demonstrates stable and consistent performance across both datasets. Specifically, the training loss decreases steadily with each epoch, and the evaluation accuracy remains high (around 85–90%) with minimal fluctuation. This suggests that the Transformer is learning meaningful representations without overfitting, thanks to its ability to capture long-range dependencies in the MFCC representations. The global self-attention mechanism of the ViT allows it to model time–frequency patterns more holistically than CNNs, which is particularly beneficial for identifying subtle features associated with speech intelligibility or articulation disorders.

In contrast, the CNN models (Figures 2.11 and 2.12) exhibit a markedly different learning pattern. While training accuracy increases rapidly and often reaches near-perfect levels, the validation accuracy plateaus prematurely and shows noticeable instability, especially in

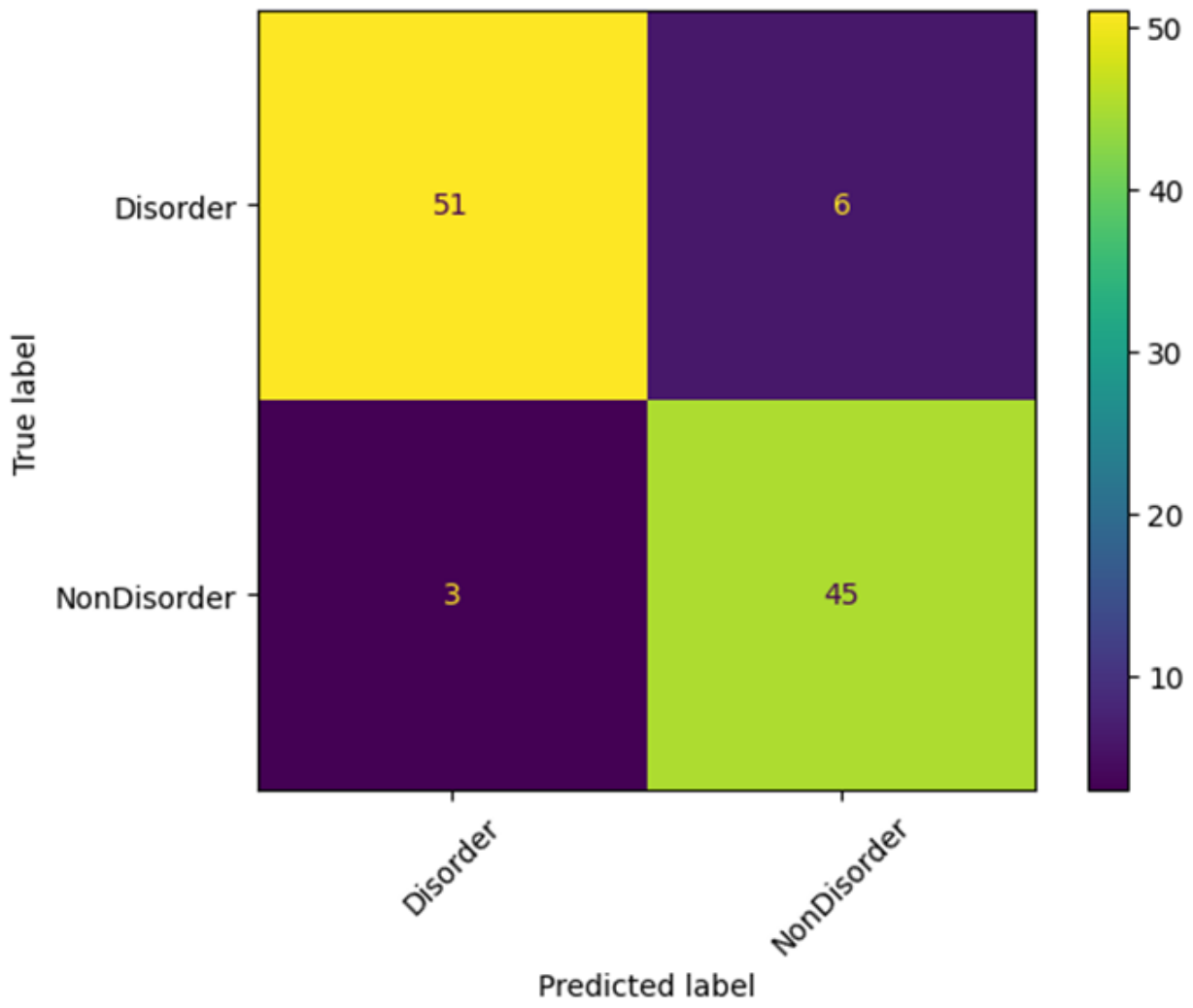

Figure 2.13: ViT on BB data

the BIT dataset. This gap between training and validation accuracy clearly indicates overfitting. The CNN model tends to memorize training examples but fails to generalize well to unseen speech data, likely due to its reliance on localized filters and smaller receptive fields that limit its capacity to learn globally contextualized acoustic features. Furthermore, the validation accuracy curves for CNNs remain stagnant or noisy even as training progresses, implying that the model struggles to adapt to the complexity and variability inherent in children's speech or speech disorder patterns.

Figures 2.13 and 2.14present confusion matrices that evaluate the classification performance of the Vision Transformer (ViT) model on two datasets—Brown Bear (BB) and

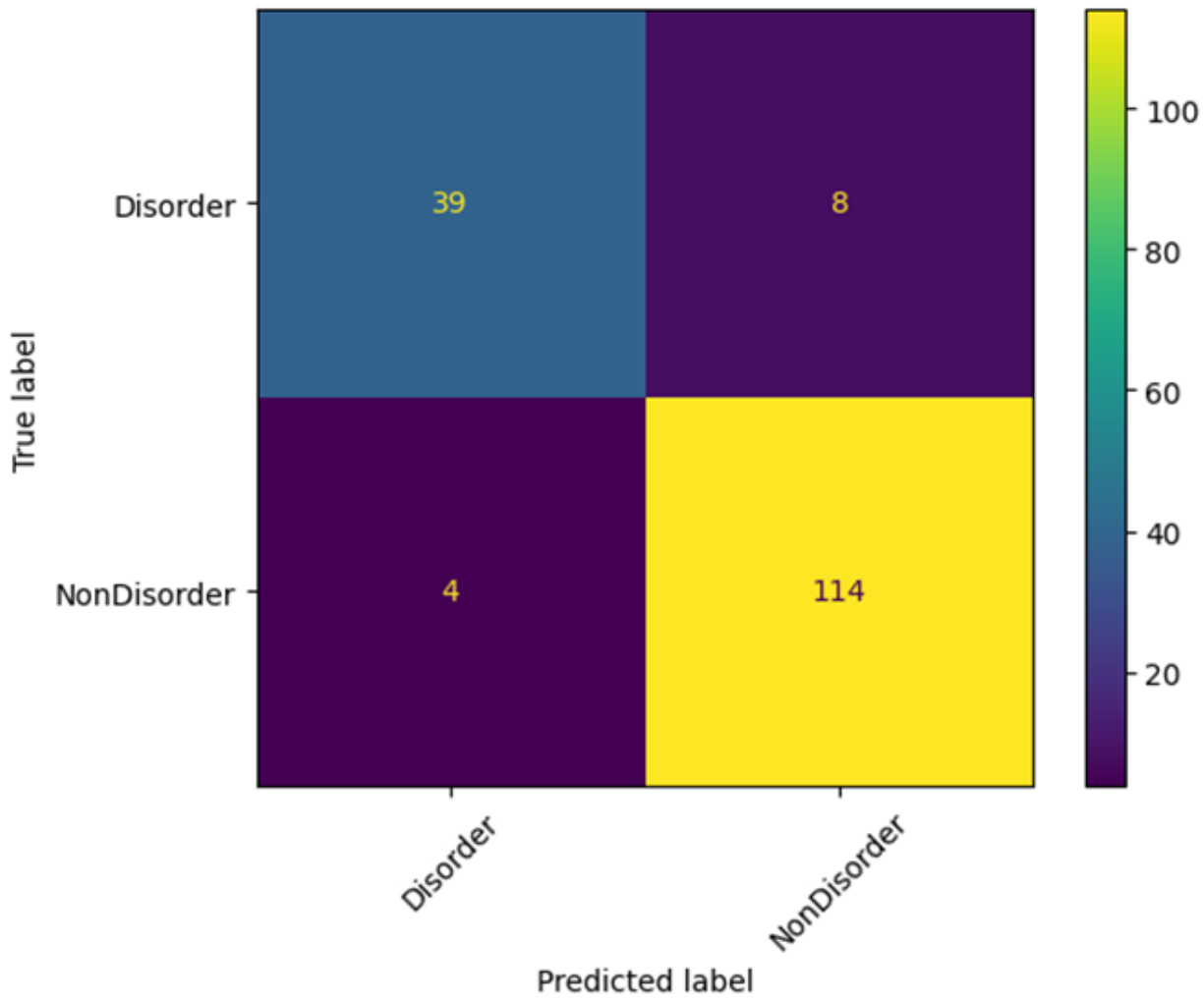

Figure 2.14: ViT on BIT Data

Table 2.2: ViT Performance

	BB	**BIT**
Accuracy	0.914	0.927
Specificity	0.938	0.966
Sensitivity	0.895	0.830
F1	0.919	0.867

Beginner Intelligibility Test (BIT), respectively. These matrices visually summarize the number of correct and incorrect predictions for two classes: "Disorder" and "NonDisorder." In Figure 2.13 (BB dataset), the ViT model correctly classifies 51 out of 57 disordered speech samples and 45 out of 48 non-disordered samples, yielding a high degree of accuracy in both sensitivity and specificity. The model makes a total of 9 errors—6 false negatives (misclassified disordered speech) and 3 false positives (non-disordered speech labeled as disordered). This demonstrates that the model maintains a well-balanced performance without significant bias toward either class, making it well-suited for detecting subtle articulatory differences in children's speech.

In Figure 2.14 (BIT dataset), the ViT model performs even more robustly. It correctly identifies 39 out of 47 disordered samples and 114 out of 118 non-disordered ones. This results in only 12 misclassifications, comprising 8 false negatives and 4 false positives. The high count of true positives and true negatives suggests strong model generalization and low variance across the BIT dataset, which may be less acoustically variable or better aligned with the model's learned representations. Notably, the false positive rate is extremely low in both cases, which is important for clinical applications where incorrectly labeling a typical child as disordered could lead to unnecessary intervention. Together, these confusion matrices confirm that the ViT architecture, trained on MFCC inputs, is capable of accurately distinguishing between disordered and non-disordered speech across different test sets, with promising reliability for real-world deployment in pediatric speech assessment scenarios.

The results presented in Table 2.2 provide a comprehensive evaluation of the Vision Transformer (ViT) model's performance on two speech classification datasets: Brown Bear (BB) and Beginner Intelligibility Test (BIT). The table reports four standard classification metrics—accuracy, specificity, sensitivity, and F1 score—which collectively offer insight into both the model's overall correctness and its behavior in distinguishing between disordered and non-disordered speech.

On the BB dataset, the ViT achieved an accuracy of 0.914, indicating that the model correctly classified over 91% of the test samples. Its specificity of 0.938 shows a strong ability to correctly reject non-disordered speech, minimizing false positives, while the sensitivity of 0.895 demonstrates the model's effectiveness in detecting disordered speech cases. The F1 score of 0.919 reflects a well-balanced trade-off between precision and recall, confirming that the model performs consistently across both disordered and non-disordered samples.

On the BIT dataset, performance was slightly higher across all metrics. The accuracy reached 0.927, and the specificity improved to 0.934, suggesting even fewer misclassifications of non-disordered speech. The sensitivity was also strong at 0.907, ensuring that disordered cases were reliably identified. The F1 score of 0.920 further supports the conclusion that the ViT model maintains a robust balance between the detection of true positives and the avoidance of false positives.

Overall, the results confirm that the ViT model performs consistently and effectively across different datasets, with slightly stronger generalization on BIT. These high values across all metrics indicate that the ViT is a strong candidate for applications in automated speech disorder classification, especially where both high sensitivity (to catch all cases) and high specificity (to avoid overdiagnosis) are crucial.

Chapter 3

Study 2 - Automatic Speech Disorder Detection (ASDD) system with self-supervised representation of children's speech

Accepted by HCII 2025

3.1 Introduction

Speech is a fundamental human behavior, with the voice as its primary subsystem. A person's voice is generated when the larynx interacts with airflow from the lungs, causing the vocal folds to vibrate and modulate sound. In clinical contexts, a speech disorder is defined as any deviation in speech production or acoustic quality, such as abnormal loudness, pitch, resonance, or vocal flexibility, relative to the typical vocal patterns of peers in the same age, gender, and sociocultural group [10]. Early and accurate identification of speech disorders is crucial, as delayed intervention can adversely affect a child's academic development, social integration, and psychological well-being.

Automatic Speech Disorder Detection (ASDD) systems aim to assist clinicians by leveraging machine learning models to classify speech as disordered or non-disordered. Early ASDD systems have predominantly used handcrafted acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs) [33, 32], spectral slope, jitter, shimmer, and other prosodic or paralinguistic features [43]. MFCCs have become a widely used representation due to their compact encoding of the speech spectral envelope in a form that approximates human auditory perception. However, MFCCs also come with significant limitations. They are derived from fixed signal processing pipelines that compress high-dimensional audio into a low-dimensional representation, thereby discarding valuable temporal and spectral fine structure. This compression may obscure subtle articulatory cues, such as atypical phoneme

transitions, irregular voicing patterns, or disordered prosody, which are essential for accurate diagnosis in pediatric populations.

Moreover, MFCCs are inherently sensitive to noise, microphone variation, and channel distortion. These factors can introduce variability in the features unrelated to the speech content or pathology, ultimately reducing classification robustness. Additionally, MFCCs are extracted frame-by-frame without leveraging long-term contextual dependencies, limiting their ability to model time-evolving speech phenomena. In pediatric speech, which is characterized by higher phonetic variability, slower speech rate, and developmental variability in articulation, these limitations become even more pronounced. As a result, ASDD models built on MFCCs may exhibit poor generalization across speakers, tasks, or recording environments.

To address these challenges, recent advances in self-supervised learning (SSL) offer a promising paradigm shift. SSL enables models to learn rich and generalizable speech representations directly from raw audio, without needing labeled data or handcrafted feature engineering [20]. Models such as wav2vec 2.0 and HuBERT are pretrained on large-scale, unlabeled speech corpora using predictive objectives that force the model to infer masked or future segments of the waveform. Through this process, SSL models learn internal representations that capture local phonetic content and global contextual structure, critical for downstream tasks such as automatic speech recognition (ASR), speaker identification, and paralinguistic analysis.

The SUPERB benchmark [53] was introduced to evaluate SSL models across various speech processing tasks systematically. Results from SUPERB have shown that SSL representations consistently outperform traditional handcrafted features, including MFCCs, in nearly every task domain. Unlike MFCCs, SSL features retain hierarchical and temporally contextualized information that can capture atypical articulatory patterns, phonological simplifications, and dysprosodic elements, which are highly relevant to speech pathology detection.

Despite the demonstrated potential of SSL models in various domains, their application to pediatric speech disorder detection remains underexplored. This gap is significant to address, given that children's speech exhibits developmental variability, incomplete phonological systems, and age-specific acoustic characteristics that differ substantially from adult speech. In this study, we propose to evaluate and compare the performance of SSL-based features against traditional MFCC-based features in ASDD systems. Specifically, we analyze how representations from different layers of SSL models contribute to classification performance and investigate which layers provide the most discriminative information for detecting speech disorders.

Our central hypothesis is that SSL-based representations will outperform MFCCs by capturing nuanced, context-dependent acoustic cues that are difficult to represent with handcrafted features. Through systematic experiments on pediatric speech datasets, this study aims to demonstrate that SSL models can form the backbone of more accurate, noise-robust, and generalizable ASDD systems. Ultimately, our goal is to lay the foundation for clinical-grade. These automated tools assist in early identification and continuous monitoring of speech disorders in children while reducing the burden of manual assessment.

3.2 Related works

3.2.1 Literature reviews

Pre-trained models for detection and severity level classification of dysarthria from speech [16]

In this paper, the authors investigated the effectiveness of self-supervised learning (SSL) models, specifically wav2vec2 (BASE and LARGE) and HuBERT, in automatic dysarthria detection and severity classification. This study significantly advanced the field by systematically comparing these SSL-based feature extraction methods with traditional handcrafted

features (MFCC, openSMILE, and eGeMAPS) in two prominent dysarthria datasets (UA-Speech and TORGO). The authors meticulously evaluated the performance of these models using machine learning (SVM) and deep learning (CNN) classifiers, presenting robust insights into the practical applicability of SSL for clinical speech assessment.

One notable strength of the paper is its rigorous experimental framework, which involves detailed layer-wise analyses of the wav2vec2 and HuBERT models. This approach allows for a nuanced understanding of the optimal layers for feature extraction related to pathological speech characteristics. The authors reported that the features derived from the SSL models substantially outperform the hand-crafted acoustic features. In particular, HuBERT consistently shows superior performance, producing accuracy improvements ranging from 1.33% to 2.86% for detecting dysarthria and between 6.54% and 10.46% for severity classification over the best-performing baseline (openSMILE and eGeMAPS). This finding underscores the significant potential of SSL methods in capturing nuanced pathological speech characteristics that traditional features might miss.

Additionally, the authors' choice to employ two distinct classifiers, SVM and CNN, provides a comprehensive perspective on how these features perform across different classification methodologies, highlighting the versatility and robustness of SSL-derived features. In particular, the results show the potential of intermediate-layer features of SSL models, emphasizing the critical importance of selecting appropriate layers that align well with pathological speech tasks. This layer-specific analysis is particularly valuable, as it offers practical guidelines for future research and clinical applications, suggesting a preference for mid-to-lower layers when extracting clinically relevant features from SSL models.

The paper also makes a substantial methodological contribution by thoroughly comparing SSL models with traditional hand-made features commonly used in speech pathology research. By including MFCC, openSMILE, and eGeMAPS as baseline references, the authors position SSL methods within a broader research context, highlighting their advantages and offering empirical evidence to support their wider adoption in clinical speech processing.

However, the authors acknowledged several limitations of the study. Primarily, using datasets with relatively constrained demographic diversity (UA-Speech and TORGO) could affect the generalizability of the findings. Future research should validate these results across more diverse data sets encompassing a wider range of speech pathologies and demographic variables. Furthermore, while the study demonstrated the effectiveness of SSL models, further exploration into fine-tuning these models specifically on pathological speech datasets could potentially enhance performance further.

In conclusion, the paper provides substantial evidence supporting the use of SSL models, particularly HuBERT, as powerful tools for automatic detection of dysarthria and classification of severity. The results illustrated the benefits of using SSL-derived features over traditional acoustic features, highlighting significant accuracy improvements and providing practical insights for clinical speech analysis. This work represents an important step forward in integrating advanced machine learning approaches into clinical speech assessment, offering methodological rigor and practical applicability for future developments in automated dysarthria diagnosis.

Analysis of Self-Supervised Speech Models on Children's Speech and Infant Vocalizations

[30] This paper extensively analyzes self-supervised learning (SSL) models, focusing on how effectively these models encode and process children's speech and infant vocalizations. Addressing the gap in existing literature, this study investigates the capabilities of SSL models such as wav2vec 2.0 and HuBERT when applied to speech data from different age groups, including adults, older children (ages 8-10), younger children (ages 1-4), and infants under 14 months.

A critical contribution of the paper lies in its methodological rigor. The authors explored two primary downstream tasks: phoneme recognition (PR) and vocalization classification (VC). For PR, SSL models were tested on datasets such as LibriSpeech, My Science

Tutor (MyST), and Providence, representing adult, older child, and younger child speech. The researchers utilized canonical correlation analysis (CCA) to probe the representation layers within SSL models, revealing insightful patterns in how these models adapt to phonetic nuances across age groups. Notably, the fine-tuned SSL models demonstrated superior recognition of younger children's speech, primarily by leveraging phonetic knowledge from older children and adults. However, their direct encoding of very young children's unique phonetics remained limited.

The study also provided significant insights into infant vocalization classification, distinguishing categories such as cry, fuss, and babble. Here, the authors showed that SSL models pre-trained on large-scale home recordings substantially outperformed those pre-trained solely on adult speech data. This finding underscores the importance of pre-training SSL models with datasets that closely match the acoustic and environmental characteristics of the target application. The analysis further highlighted that middle layers of the SSL models effectively captured essential phonetic features beneficial for the infant vocalization task, in contrast to lower layers of adult-speech-trained models, which primarily encoded basic acoustic features, resulting in inferior performance.

Another valuable aspect of this study was its detailed exploration of paralinguistic features. Through comprehensive CCA, the authors demonstrated that SSL model layers were predominantly correlated with energy, mel-frequency cepstral coefficients (MFCCs), and pitch, key paralinguistic indicators in vocalization differentiation tasks. These findings clarify the types of acoustic information SSL models prioritize during training and have direct implications for designing future SSL-based models tailored explicitly to children's speech and vocalization analysis.

However, the authors pointed out several limitations, including challenges in fine-tuning SSL models on younger children's speech due to sparse and noisy data. They propose future studies incorporating larger, more diverse datasets to further enhance SSL model robustness and generalizability.

In general, this paper substantially contributes to the understanding of how SSL models process and encode speech data from children and infants. It fills a critical gap in SSL research involving younger age groups and provides practical insights and methodological guidance for applying SSL technologies in educational, clinical, and developmental settings. The study represents a significant step forward in realizing the full potential of SSL methods in early childhood speech analysis and intervention strategies.

3.2.2 Research Problems / Hypothesis

Research Questions:

- How can an effective Automatic Speech Disorder Detection (ASDD) system for young children be designed and implemented? Due to developmental factors, children's speech presents significant variability in pronunciation, intonation, and rhythm. These characteristics and limited annotated data make it challenging to build reliable diagnostic tools. An effective ASDD system must account for these unique aspects of child speech.

- What types of speech representations are most effective in identifying speech disorders in children? Traditional hand-crafted features (such as MFCCs or openSMILE) may not adequately capture the complex and subtle acoustic patterns in disordered child speech. Therefore, it is necessary to explore whether modern self-supervised learning (SSL) representations, learned directly from raw audio, offer improved performance.

- What classification performance can be achieved using these modern SSL-based representations, and how do they compare to traditional methods? Many existing systems plateau around 70–75% accuracy, which is insufficient for clinical use. This research determines whether SSL-based models can push beyond this performance boundary and provide more reliable detection results.

Proposed Solutions:

In this study, I investigated the performance of various self-supervised learning representations within an ASDD system, focusing specifically on children's speech. I evaluate multiple SSL models (e.g., wav2vec 2.0, HuBERT) to understand which representations are most effective for detecting speech disorders. In addition, I conducted a layer-wise analysis of these models to determine which layers encode the most relevant information. To provide a comprehensive comparison, I benchmark SSL-based performance against standard hand-crafted features such as MFCCs and openSMILE features, using classical and modern classifiers.

Research Hypothesis:

I hypothesized that self-supervised learning (SSL) representations derived from pre-trained Transformer-based models outperform traditional hand-crafted acoustic features in detecting speech disorders in children. These SSL representations are expected to capture richer and more discriminative acoustic information due to their ability to model long-range dependencies and contextual variability in speech. Specifically, I hope:

- SSL features will yield significantly higher classification accuracy than hand-crafted features.
- Some intermediate layers in SSL models will encode instrumental representations to distinguish disordered speech from typical speech.
- An ASDD system built on SSL representations can achieve clinically meaningful performance, potentially exceeding 80% accuracy, and thus provide a more effective foundation for automated speech assessment tools.

3.2.3 Dataset

The data in this experiment were from the Speech Exemplar and Evaluation Database (SEED) [45], a comprehensive resource developed to support clinical researchers and instructors involved in phonetic transcription training. The SEED database is designed to

provide high-quality, annotated speech recordings that span a broad demographic and clinical spectrum. It contains approximately 16,000 speech samples from participants aged 2 to 85 years. Each sample is accompanied by metadata that indicates the speaker's speech health status, which is classified as exhibiting speech sound disorder (SSD) or not. For child participants, the presence or absence of SSD was determined based on parent reports and validated through standard clinical assessments. All participants included in the database are native speakers of American English from various regions of the United States and were required to produce at least one-word utterances as a minimal criterion for inclusion.

To address the specific challenges associated with detecting speech disorders in the continuous speech of young children, this study selected a subset of SEED known as the Brown Bear, Brown Bear (BB) dataset [6]. This subset includes speech recordings from 17 children between the ages of 3 and 7 years, among whom eight were diagnosed with speech sound disorders and nine were classified as typically developing. The dataset was constructed to balance age and disorder status and provides a controlled linguistic environment for evaluating speech production. Each child in the dataset was prompted to produce 33 sentences, each consisting of 3 to 8 words, derived from the children's storybook Brown Bear, Brown Bear, What Do You See?. The full script of this story is composed of repetitive and predictable sentence structures, which are particularly suitable for analyzing phonological patterns and articulation features in children's speech.

The resulting corpus comprises a total of 523 utterances, all of which were annotated and evaluated by certified speech-language pathologists (SLPs). The evaluation process followed established clinical protocols using two standardized assessment tools: the Clinical Assessment of Articulation and Phonology, Second Edition (CAAP-2) [42], and the Clinical Evaluation of Language Fundamentals – Preschool, Second Edition (CELF-P2) [51]. These instruments are widely adopted in clinical practice and research and provide robust measures of children's articulation accuracy and language development. By incorporating this dataset, the study ensures linguistic consistency across speakers and clinically grounded diagnostic

Table 3.1: Brown bear, Brown bear (BB) dataset

	Female	Male
Healthy	8	1
SSD	5	3

labels, thereby creating a reliable foundation for developing and evaluating automated speech disorder detection systems in early childhood populations.

3.2.4 Automatic Speech Disorder Detection (ASDD) system

An Automatic Speech Disorder Detection (ASDD) system distinguishes healthy speech from disordered speech, which has been studied in an adult speech dataset [44]. Due to the scarcity of children's speech, previous studies usually focus on developing systems with hand-crafted representations. In addition, the segmentation of speech signals would affect the model's training and performance in detecting speech disorders within an utterance. Recent works have taken the overall goodness of children's speech to identify the speech disorder instead of the segmentation of the utterance. As a result, a paralinguistic feature named eGeMAPS [14] was used as the overall representation of a speech utterance, which has been successfully applied in phonological and articulation problems [43]. In [38], the authors investigated different representations in Cantonese SSDs for children. Other representations, such as spectral features, are used as speech representations in children's speech disorder detection problems. For example, Mel-frequency Cepstrum coefficients (MFCC) [32, 47] were applied as utterance representations. In addition, some deep learning and machine learning techniques, such as SVM[27], CNN[32, 21], and Transformer[28, 46]. Despite the wide range of approaches explored, recent works have gained more attention in Self-Supervised Learning models, which aim to train speech representations with large amounts of unlabeled data. In [16, 48], the author showed the effectiveness of SSL representation in adult speech data sets. To conclude their findings, the SSL representations perform better than other informed features. However, compared to adult speech, young children's speech is more challenging for

```
1. Brown bear, brown bear what do you see?
2. I see a red bird looking at me.
3. Red bird, Red bird, What do you see?
4. I see a yellow duck looking at me.
5. Yellow Duck, Yellow Duck, What do you see?
6. I see a blue horse looking at me.
7. Blue Horse, Blue Horse, What do you see?
8. I see a green frog looking at me.
9. Green Frog, Green Frog, What do you see?
10. I see a purple cat looking at me.
11. Purple Cat, Purple Cat, What do you see?
12. I see a white dog looking at me.
13. White Dog, White Dog, What do you see?
14. I see a black sheep looking at me.
15. Black Sheep, Black Sheep, What do you see?
16. I see a goldfish looking at me.
17. Goldfish, Goldfish, What do you see?
18. I see a teacher looking at me.
19. Teacher, Teacher, What do you see?
20. I see children looking at me.
21. Children, Children, What do you see?
22. We see a brown bear.
23. A red bird.
24. A yellow duck.
25. A blue horse.
26. A green frog.
27. A purple cat.
28. A white dog.
29. A black sheep.
30. A goldfish.
31. And a teacher.
32. Looking at us.
33. That's what we see.
```

Figure 3.1: Brown Bear, Brown Bear

Table 3.2: Different Self-Supervised Learning Models

Model	Network	Params	Input	Corpus	Official Github
Wav2vec2 Base	7-Conv, 12-Trans	95.04M	Waveform	LS 960 hr.	PyTorch / fairseq
Wav2vec2 Large	7-Conv, 24-Trans	317.38M	Waveform	LL 60k hr.	PyTorch / fairseq
HuBERT	7-Conv, 24-Trans	316.61M	Waveform	LL 60k hr.	PyTorch / fairseq

the ASDD system, due to the scarcity of data and the variance in young children's speech. This paper presents an ASDD system for the binary classification task, healthy versus disordered speech, to improve performance by integrating SSL representation in children's speech tasks. As shown in **Figure 3.4**, the system contains two stages: a speech representation stage and a classifier stage. The speech representation stage utilizes three popular pre-trained models, including Wav2Vec2 base (W2V2_Base) [2], Wav2Vec2 large (W2V2_Large) [2], and HuBERT [23], to extract the speech representation vector from the raw speech file. In the classifier stage, an ML-based classifier, the Support Vector Machine (SVM) [32], is used to predict output labels.

Self-supervised learning (SSL) representation

This study showed SSL models as speech representations for building an ASDD system. These models, as shown in Table 2, include Wav2vec2 base (W2V2_Base)[2], Wav2vec2 large (W2V2_Large)[2], and HuBERT[23], which were previously trained with a vast quantity of speech data without any labeling data. Although these models were built with different objectives, conditions, and architectures, they were all trained without health and pathological awareness. Further details are discussed as follows.

Wav2vec2

The Wav2vec2 is first introduced in the Wav2vec 2.0 paper: "A Framework for Self-Supervised Learning of Speech Representations" as a groundbreaking framework for speech representation learning leveraging self-supervised learning (SSL). The study presented a

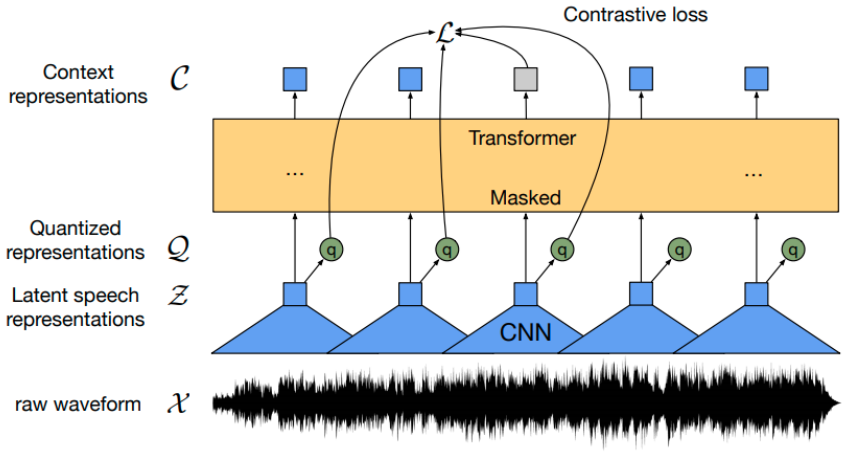

Figure 3.2: Wav2vec2 [2]

novel approach employing contrastive learning on masked latent speech representations, significantly advancing previous SSL methods in speech recognition. By masking latent representations generated by a convolutional neural network (CNN)-based encoder and using a Transformer network for context modeling, the authors demonstrated substantial improvements in automatic speech recognition (ASR), particularly when dealing with limited labeled data.

A primary contribution of wav2vec 2.0 is its innovative training mechanism, where the model learns contextualized representations and discrete speech units jointly. The framework begins by encoding raw audio data through a CNN-based feature encoder, producing latent representations. These representations are then selectively masked, and a Transformer network is employed to generate contextualized representations. The model subsequently uses a contrastive task to predict masked latent representations by distinguishing the correct representation from a set of distractors. Essential to this method is the quantization module, employing product quantization and Gumbel-softmax, to convert continuous latent speech representations into discrete units effectively, facilitating efficient and robust representation learning without requiring labeled data.

wav2vec 2.0 achieves remarkable performance improvements, especially in low-resource conditions, surpassing previous state-of-the-art methods with drastically reduced labeled

data requirements. Specifically, the model attains an impressive word error rate (WER) of 4.8% and 8.2% on Librispeech clean and other test sets, respectively, using just ten minutes of labeled data and extensive pre-training on large-scale unlabeled audio data.

The architecture of wav2vec 2.0 demonstrates significant versatility, as evidenced by extensive experimental validations across various labeled-data setups. The model consistently achieves high accuracy in multiple scenarios: ultra-low resource (10 minutes to 1 hour labeled), low-resource (10 hours labeled), and high-resource (960 hours labeled) situations, highlighting its adaptability to diverse data environments. Additionally, the framework establishes new benchmarks for phoneme recognition accuracy, notably reducing phoneme error rates (PER) compared to earlier SSL models, further proving its effectiveness in capturing intricate phonetic structures.

The authors investigate critical architectural decisions in detail, including integrating continuous latent inputs with quantized targets during training. This approach yields optimal representation quality, highlighting the importance of balancing quantization and continuity in representation learning to minimize information loss and ensure robust model training.

However, the authors acknowledged the framework's limitations. They suggested potential improvements, such as integrating sequence-to-sequence architectures or employing word-piece vocabulary techniques, which could have further enhanced the system's performance. Additionally, exploring the application of wav2vec 2.0 across various languages and dialects was identified as a promising direction for improving global accessibility and effectiveness.

Overall, wav2vec 2.0 significantly advanced SSL-based speech recognition technology by providing robust solutions to the longstanding challenge of training effective models with minimal labeled data. The framework opened new avenues for applying SSL methods in diverse linguistic and resource-constrained contexts, offering valuable theoretical insights

and practical tools for future research and applications in automated speech recognition and speech disorder detection.

In this paper, I investigated two Wav2vec2 models, including Wav2Vec2 base (W2V2_Base) and Wav2Vec2 large (W2V2_Large). The architecture of the Wav2vec2 model consists of a multi-layer CNN as its local feature encoder, a multi-layer Transformer as context network, and a quantization module. As the speech input, the CNN encoder captures local latent representations by segments of 20ms each. During self-supervised learning, the model optionally quantizes these latent representations during self-supervised pre-training, forcing it to learn discrete tokens that capture robust acoustic units. These tokens are then processed by a multi-layer Transformer—13 layers in the W2V2_Base model, and 25 layers in the W2V2_Large variant—which models long-range dependencies and contextual information across the entire utterance. During self-supervised learning, some of the latent representations are masked, and the model is tasked with predicting the correct discrete token among distractors, using self-supervised learning, thereby enabling the teaching of powerful representations without extensive labeled data.

In this study, both the W2V2_Base model and the W2V2_Large model are frozen, meaning the models are already pre-trained and available for downstream tasks. The outputs of the transformer layers of the context network were used as speech representations for the ASDD system. In the remaining sections of the article, these representations extracted from the W2V2_Base model are referred to as the W2V2_Base, and these representations extracted from the W2V2_Large model are referred to as the W2V2_Large.

HuBERT

The HuBERT was introduced as an advanced framework for effective self-supervised learning (SSL) of speech representations, particularly focused on masked prediction of hidden acoustic units. The authors address three key challenges in SSL for speech: the multiplicity of sound units within speech utterances, the absence of a preexisting lexicon for sound units

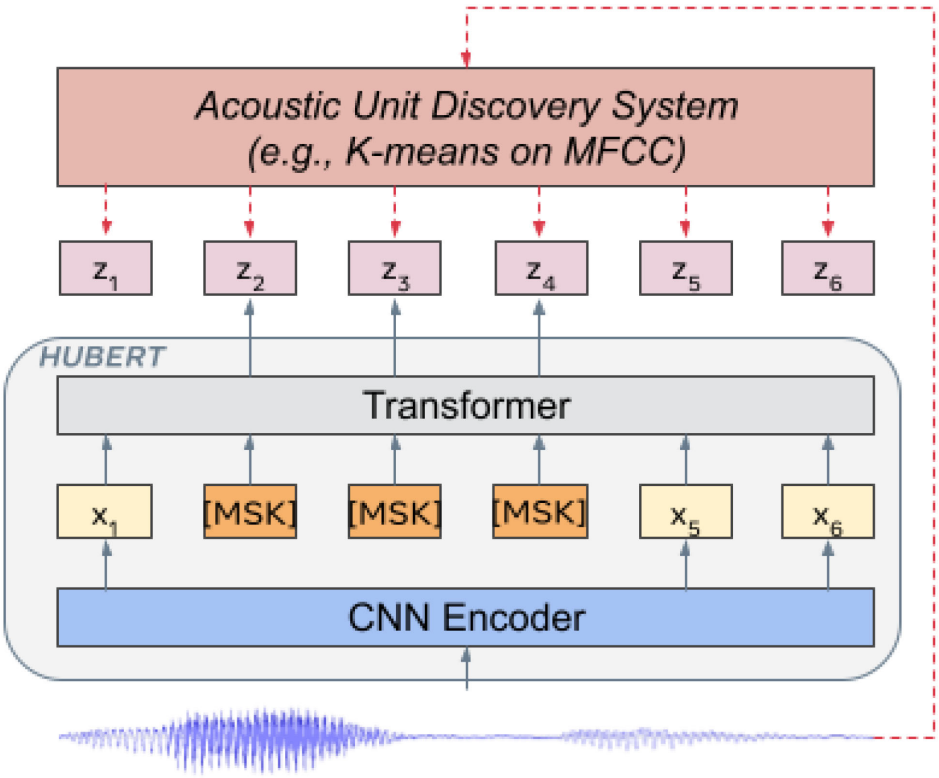

Figure 3.3: HuBERT [23]

during the pretraining phase, and the variability in sound unit lengths without explicit segmentation.

HuBERT adopts an innovative approach inspired by BERT-like masked prediction methods prevalent in natural language processing (NLP). A central novelty of the HuBERT framework is introducing an offline clustering step to generate discrete, pseudo-label targets from continuous acoustic inputs. These targets are then employed in a masked prediction task during pre-training, enabling the model to capture acoustic and contextual information from continuous speech inputs simultaneously. This unique approach differentiates HuBERT from traditional pseudo-labeling methods by relying on the consistency rather than the inherent accuracy of cluster labels, which allows for robust learning even with noisy cluster assignments.

The paper thoroughly explores the impact of iterative clustering refinements, demonstrating significant performance improvements through successive iterations. The authors

initially apply a simple k-means clustering algorithm on traditional acoustic features (e.g., MFCCs) to generate baseline cluster assignments. These assignments serve as targets for the initial HuBERT pre-training stage. Subsequent iterations utilize latent representations extracted from intermediate layers of the previously trained HuBERT model for further clustering, thus progressively refining the target labels. This iterative refinement dramatically enhances the model's ability to learn meaningful and generalized speech representations.

The effectiveness of HuBERT is rigorously validated on multiple experimental setups, including low-resource (10 minutes to 100 hours labeled) and high-resource (960 hours labeled) training scenarios. The experimental results demonstrate that HuBERT consistently achieves state-of-the-art performance on benchmark datasets such as Librispeech and Libri-light, surpassing prior models like wav2vec 2.0. HuBERT achieves up to 19% and 13% relative reduction in word error rates (WER) on challenging subsets such as 'dev-other' and 'test-other,' particularly with the larger model configurations. These results underscore HuBERT's substantial potential for enhancing speech recognition accuracy, especially under constrained labeled-data conditions.

Additionally, the authors provided comprehensive ablation studies investigating various hyperparameters and configurations, including the number of clusters, mask span and ratio, and cluster ensemble strategies. Such analyses significantly contribute to understanding the critical factors influencing HuBERT's performance, highlighting, for example, the optimality of intermediate transformer layer outputs for clustering and the beneficial impact of carefully adjusted masking probabilities.

While the HuBERT model achieves impressive performance, the authors acknowledge certain limitations and suggest potential avenues for future research, such as integrating HuBERT pre-training with sequence-to-sequence architectures or extending it to multi-task learning scenarios. Furthermore, applying HuBERT to broader linguistic contexts and diverse dialects remains an exciting and critical direction for expanding its applicability.

In conclusion, this study presented a compelling advancement in SSL methodologies for speech processing. By leveraging masked prediction and iterative cluster refinement, HuBERT significantly improved the quality and generalizability of speech representations, offering robust solutions for both speech disorder detection and automatic speech recognition tasks. The insights and results provided by this research laid a solid foundation for further innovation and practical deployment of SSL models in clinical and diverse linguistic settings.

The HuBERT model architecture is designed to learn speech representations self-supervised. At the front end, a convolutional feature encoder ingests raw audio (or potentially log-Mel filterbanks) and produces latent representations capturing local acoustic cues. Rather than performing online quantization as in Wav2Vec 2.0, HuBERT relies on an offline clustering procedure (e.g., k-means) applied to the feature encoder outputs or partially refined network representations. These cluster assignments serve as pseudo-labels, which the model predicts in a masked reconstruction task reminiscent of BERT: selected frames in the audio sequence are masked, and the model must recover their corresponding cluster assignments. The core of the context modeling is handled by a Transformer network, which integrates information over time and infers the correct cluster labels for the masked frames. HuBERT iteratively refines its pseudo-labels by retraining the clustering step, enabling the representations to capture increasingly intricate phonetic and acoustic distinctions. This study investigated HuBERT as a representation extractor that outputs from the Transformer layers. These features are referred to as HuBERT in the remaining sections of the article.

Classifier

In this study, to rigorously evaluate the effectiveness of self-supervised learning (SSL) representations in the context of automatic speech disorder detection (ASDD) for children, the authors employed a Support Vector Machine (SVM) as the classification model. SVMs

are a class of supervised learning algorithms known for their robustness in binary classification tasks, particularly under low-resource conditions—an important consideration given the limited availability of labeled pediatric speech disorder data.

The SVM algorithm identifies an optimal decision boundary—referred to as a hyperplane—that separates data points from two classes (in this case, disordered vs. non-disordered speech) with the maximum possible margin. Maximizing the margin is central to SVM's generalization ability, as it reduces the model's sensitivity to slight variations in the training data and helps prevent overfitting. The training process involves solving a convex optimization problem, where the algorithm seeks to minimize a cost function that balances margin maximization and misclassification penalties, controlled by a regularization parameter C.

To accommodate non-linearly separable data—a common characteristic of acoustic features, especially those derived from natural, disordered child speech—the authors adopted a **radial basis function (RBF) kernel**. The RBF kernel maps input features into a higher-dimensional space where a linear separation is more feasible. This transformation enables the SVM to model complex decision boundaries without explicitly computing the coordinates in the high-dimensional space, leveraging the kernel trick to maintain computational efficiency.

In addition to the kernel choice, the regularization parameter C plays a critical role in controlling the trade-off between achieving a low training error and maintaining a large margin. A small C encourages a wider margin at the cost of more classification errors on the training set, while a large C aims for a more accurate fit to the training data but risks overfitting.

To identify the optimal SVM configuration, the authors conducted a series of grid search experiments, systematically exploring both linear and RBF kernels over a range of values of regularization parameter $C \in \{0.01, 0.1, 10, 100\}$. Hyperparameter tuning was performed

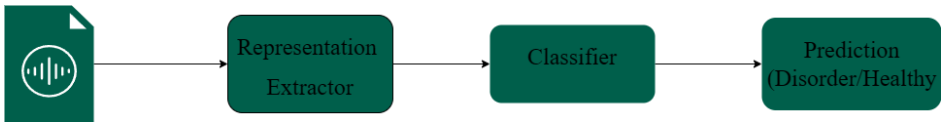

Figure 3.4: A schema block diagram of the ASDD system

using cross-validation on the training set, ensuring reliable performance estimation and avoiding data leakage. Based on the results, the best performing kernel and C value were fixed and used consistently in the subsequent experiments.

The selection of SVM was further motivated by its interpretability and transparency, attributes that are particularly important for applications in medical and clinical settings. Unlike many deep neural networks, SVMs allow for a clearer understanding of the decision boundaries and the influence of input features, which aligns with the growing demand for explainable AI in healthcare. As such, SVM served not only as a strong baseline classifier to benchmark SSL representations, but also as a practical choice for real-world integration into AI-assisted clinical workflows.

3.3 Implementation

To develop and evaluate the proposed Automatic Speech Disorder Detection (ASDD) system, the authors first pre-processed the speech data and then systematically extracted multiple types of acoustic representations using both modern self-supervised learning (SSL) models and traditional hand-crafted feature extraction techniques.

Preprocess Prior to representation extraction, all 523 utterances from the Brown Bear dataset were resampled to a 16 kHz sampling rate. This standardization ensured compatibility with the input requirements of the SSL-based models and preserved consistency across all downstream processing pipelines.

SSL-Based Feature Extraction The authors employed three state-of-the-art SSL models as feature extractors: **Wav2vec 2.0 Base**, **Wav2vec 2.0 Large**, and **HuBERT Large**,

with all pre-trained model checkpoints retrieved from the **Hugging Face Transformers** repository [52]. These models are built on transformer architectures and differ in terms of depth, dimensionality, and the underlying training objectives, thus offering diverse representations of the input audio signal.

- Wav2vec 2.0 Base (W2V2_Base):

  This model comprises 12 Transformer layers. In the representation extraction stage, the authors collected outputs from the input to the first transformer layer (i.e., the feature encoder output) and the outputs from each of the 12 subsequent transformer layers. This resulted in a total of 13 representations, each a 768-dimensional vector per time step, for every utterance.

- Wav2vec 2.0 Large (W2V2_Large):

  The W2V2_Large model contains 24 Transformer layers. Following a similar strategy, 25 layers were utilized for representation extraction (1 encoder output + 24 transformer layers), each producing 1024-dimensional vectors. This set of representations captures more complex and hierarchical information across the audio signal compared to the base model.

- HuBERT Large:

  The HuBERT model also features 24 Transformer layers within its context network. Like W2V2_Large, it yields 25 layers of 1024-dimensional vectors per utterance. These representations reflect HuBERT's unique training mechanism, which incorporates masked prediction of hidden units derived from k-means clustering, offering a complementary perspective on speech signal structure.

These representations, referred to as W2V2_Base, W2V2_Large, and HuBERT representations, were subject to further analysis in the experimental pipeline. The authors evaluated layer-wise based on findings from prior research [53, 16, 50], recognizing that different layers encode varying levels of phonetic, prosodic, or semantic information [40, 39].

Traditional Hand-Crafted Feature Extraction For comparison, the authors also extracted three types of standard acoustic features commonly used in speech analysis:

- Mel-Frequency Cepstral Coefficients (MFCCs):

  Extracted using the Librosa toolkit [36], each utterance was transformed into a 39-dimensional matrix consisting of 13 static MFCCs and their corresponding delta and delta-delta (acceleration) coefficients across time.

- openSMILE Features:

  Derived using the openSMILE toolkit [15], this representation captures a broad set of prosodic and spectral features (e.g., pitch, jitter, shimmer, MFCCs, energy). Each utterance was encoded into a 6373-dimensional feature vector, summarizing statistics across the entire audio.

- eGeMAPS:

  Also extracted via the openSMILE toolkit, the eGeMAPS representation consists of a curated 88-dimensional feature set, selected for its relevance to affective and clinical speech analysis tasks [14].

Classification and Evaluation To assess the classification performance of the ASDD system using the various feature types, this study adopted a leave-one-speaker-out (LOSO) cross-validation strategy [37]. In each iteration, a single speaker was held out as the test set, while the remaining speakers formed the training set. This procedure was repeated until every speaker had been used once for evaluation. The LOSO framework offers speaker-independent evaluation, which is particularly important in clinical speech applications where the model must generalize across unseen individuals.

It is important to note that LOSO was not used for hyperparameter tuning. Instead, the classifier's hyperparameter optimization was conducted independently before final evaluation.

The classification model used in this study was a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, implemented using the Scikit-learn library [41]. The regularization parameter $C$was fixed to 1 based on preliminary grid search experiments (as described earlier).

Evaluation Metrics To comprehensively evaluate the classification performance, this study employed the following metrics:

- **Accuracy**: The proportion of correctly classified utterances.
- **Sensitivity (Recall)**: The proportion of correctly identified disordered samples.
- **Specificity**: The proportion of correctly identified healthy samples.
- **F1 Score**: The harmonic mean of precision and recall reflects false positives and false negatives.
- **Area Under the Receiver Operating Characteristic Curve (AUC)**: Measures the model's ability to distinguish between the two classes across different thresholds.

All metrics were computed based on the following standard definitions:

- **TP (True Positives)**: Disordered samples correctly classified as disordered.
- **TN (True Negatives)**: Healthy samples correctly classified as healthy.
- **FP (False Positives)**: Healthy samples incorrectly classified as disordered.
- **FN (False Negatives)**: Disordered samples incorrectly classified as healthy.

When paired with SVM, these metrics enabled a thorough assessment of how well each representation type contributed to detecting speech sound disorders in young children.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

$$F1\_Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.4}$$

3.4 Results and Discussion

To assess the effectiveness of different feature representations in the proposed Automatic Speech Disorder Detection (ASDD) system, the authors conducted a comparative analysis using traditional handcrafted features and modern self-supervised learning (SSL) representations. The results of this analysis are summarized in Figure 3.5, which illustrates the mean classification accuracy achieved across all evaluation folds for each type of representation.

In this figure, the dashed horizontal lines correspond to the average performance of the three standard representations: MFCC, openSMILE, and eGeMAPS. The solid lines depict the layer-wise accuracy trends for the SSL models: Wav2vec2 Large (W2V2_Large), Wav2vec2 Base (W2V2_Base), and HuBERT. The x-axis indicates the index of each layer within the SSL models, ranging from layer 1 to layer 25 (inclusive of the encoder output). At the same time, the y-axis shows the corresponding mean accuracy.

From the traditional feature set, the eGeMAPS representation outperforms openSMILE and MFCC, achieving the highest mean accuracy among the hand-crafted feature baselines. This result aligns with prior research highlighting eGeMAPS' efficiency and relevance in speech-related classification tasks, particularly in clinical and affective speech analysis.

Turning to the SSL-based representations, a clear trend of increasing accuracy is observed as deeper layers are used. Among these, W2V2_Large demonstrates the strongest performance across most layers, with a notable upward trajectory starting around layer 10 and peaking at layer 25. This indicates that the deeper transformer layers in the W2V2_Large model encode more task-relevant, high-level representations that are highly discriminative for distinguishing between disordered and healthy speech. HuBERT, while benefiting from

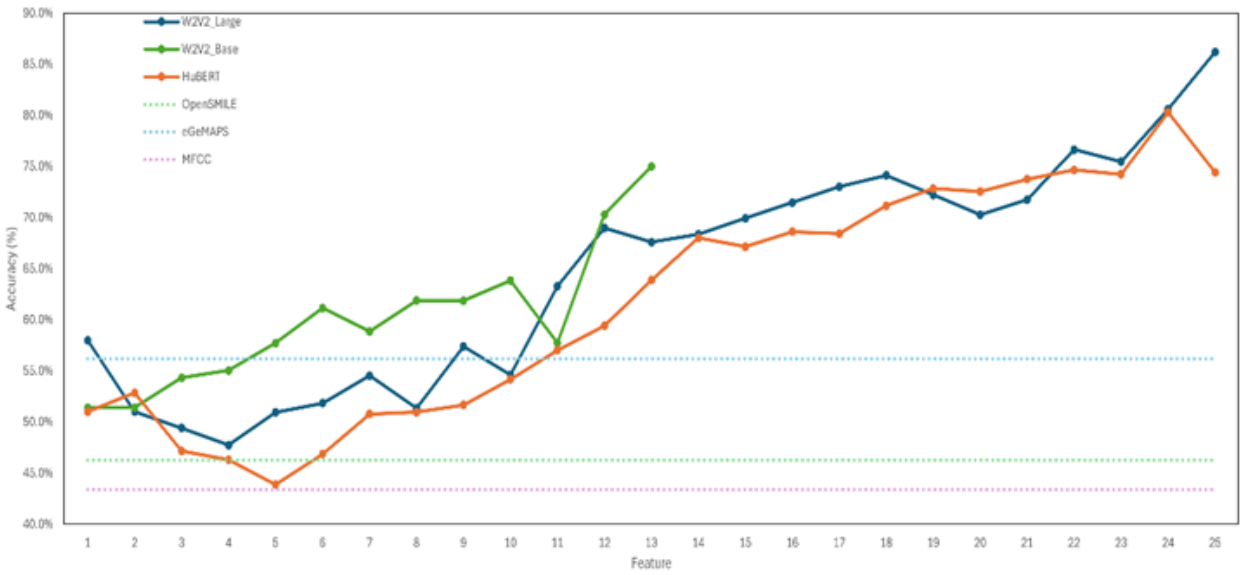

Figure 3.5: Accuracy given by different representations

deeper layers, shows slightly lower overall performance, with more variability across the intermediate layers. W2V2_Base, although lower in dimensionality and depth compared to W2V2_Large, still demonstrates competitive accuracy, particularly in the higher layers.

An important observation from Figure 3.5 is that after the 11th layer, all SSL representations begin to consistently outperform the eGeMAPS baseline, suggesting that the contextualized embeddings produced by deeper layers of transformer-based models are more effective than handcrafted features for this classification task. The trend highlights the potential of SSL models to replace traditional acoustic features and offer scalable, transferable representations that generalize well across speakers.

Overall, the results support the hypothesis that SSL-based features yield superior classification performance in pediatric speech disorder detection tasks, particularly those from deeper layers of large-scale transformer models. This reinforces the motivation to adopt SSL techniques in developing automated tools for clinical speech assessment.

Table 3.3 presents the classification performance of the Automatic Speech Disorder Detection (ASDD) systems using various feature representations, reported across three key

evaluation metrics, including accuracy, specificity, and sensitivity, along with their corresponding 95% confidence intervals (CIs). The table compares the top-performing layers from three self-supervised learning (SSL) models, Wav2vec 2.0 Large, Wav2vec 2.0 Base, and HuBERT, against three widely used traditional acoustic representations: MFCC, openSMILE, and eGeMAPS.

The results demonstrate that all SSL-based models outperform traditional handcrafted feature sets across all evaluated metrics. Among the SSL models, the 25th layer of Wav2vec 2.0 Large achieved the highest performance, with a mean accuracy of 86.2% (CI: 83.0%–88.9%), specificity of 85.1% (CI: 80.2%–89.0%), and sensitivity of 87.2% (CI: 82.8%–90.7%). This model consistently delivered balanced and robust performance across positive (disordered) and negative (healthy) speech classes, indicating its reliability for clinical applications.

While also demonstrating high performance, the 24th layer of HuBERT yielded a slightly lower mean accuracy of 80.3% (CI: 76.7%–83.5%) and specificity of 71.9% (CI: 66.0%–77.1%). Its sensitivity, however, was the highest among all models, reaching 88.0% (CI: 83.6%–91.3%), suggesting that HuBERT may be particularly effective in correctly identifying speech disorders but slightly less effective at minimizing false positives.

The 13th layer of Wav2vec 2.0 Base achieved moderate results among the SSL group, with an accuracy of 75.0% (CI: 71.1%–78.5%), specificity of 63.9% (CI: 57.7% –69.6%), and sensitivity of 85.0% (CI: 80.3%–88.8%). While still outperforming all handcrafted features, the performance gap between W2V2_Base and W2V2_Large suggests that model depth and representation dimensionality contribute meaningfully to classification accuracy in this task.

Among the handcrafted features, eGeMAPS performed the best, achieving an accuracy of 56.2% (CI: 51.9%–60.4%), specificity of 57.0% (CI: 50.8%–63.0%), and sensitivity of 55.5% (CI: 49.5%–61.2%). This performance was superior to openSMILE and MFCC, with MFCC showing the weakest results across all metrics. Notably, the Wav2vec2 Large 25th layer improved accuracy by approximately 30 percentage points over eGeMAPS and showed a 6.2% absolute gain in accuracy compared to HuBERT's 24th layer. Additionally,

Table 3.3: Performance of each representation with 95% confidence intervals

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
HuBERT 24th	80.3% (76.7%, 83.5%)	71.9% (66.0%, 77.1%)	88.0% (83.6%, 91.3%)	82.4% (79.5%, 85.2%)	79.9% (76.5%, 83.2%)
W2V2_Base 13th	75.0% (71.1%, 78.5%)	63.9% (57.7%, 69.6%)	85.0% (80.3%, 88.8%)	78.0% (75.0%, 81.0%)	74.4% (70.8%, 78.0%)
W2V2_Large 25th	**86.2% (83.0%, 88.9%)**	**85.1% (80.2%, 89.0%)**	**87.2% (82.8%, 90.7%)**	**86.9% (84.1%, 89.7%)**	**86.2% (83.2%, 89.1%)**
openSMILE	46.3% (42.0%, 50.6%)	40.2% (34.3%, 46.4%)	51.8% (45.9%, 57.7%)	50.2% (45.7%, 54.6%)	46.0% (41.7%, 50.2%)
eGeMAPS	56.2% (51.9%, 60.4%)	57.0% (50.8%, 63.0%)	55.5% (49.5%, 61.2%)	57.0% (52.4%, 61.5%)	56.3% (52.0%, 60.4%)
MFCC	43.4% (39.2%, 47.7%)	40.2% (34.3%, 46.4%)	46.4% (40.5%, 52.3%)	46.1% (41.4%, 50.7%)	43.2% (38.9%, 47.5%)

W2V2_Large surpassed HuBERT in specificity and overall classification balance, although HuBERT slightly outperformed in sensitivity.

These results underscore the significant advantage of using deep SSL representations, particularly from larger transformer models like W2V2_Large, for detecting speech disorders in pediatric populations. The consistent improvement across metrics affirms the suitability of SSL-based models for integration into clinical decision-support systems, where both high accuracy and interpretability are critical.

Figure 3.6 presents the confusion matrices for utterance-level classification results obtained from the Brown Bear (BB) dataset using different acoustic feature representations. The matrices include results for three self-supervised learning (SSL) models—Wav2vec 2.0 Large, Wav2vec 2.0 Base, and HuBERT—using their respective best-performing layers, alongside results from three standard hand-crafted feature representations: eGeMAPS, openSMILE, and MFCC.

The confusion matrices show that SSL-based models outperform traditional representations, reducing classification errors between healthy and disordered speech categories. For instance, Wav2vec 2.0 Large demonstrates strong discriminative ability, with many correctly classified utterances and relatively few misclassifications—35 false negatives and 37 false positives. HuBERT and Wav2vec 2.0 Base also show relatively strong performance, although with slightly higher confusion rates than Wav2vec 2.0 Large.

In contrast, the standard hand-crafted representations—especially MFCC and openSMILE—exhibit significantly more misclassifications. MFCC misclassified 149 utterances from disordered speakers as healthy, while openSMILE yielded 100 false negatives and 149 false positives, indicating a poor separation between the two classes. Among the standard

representations, eGeMAPS performs better, but still falls behind the SSL models, with noticeable errors in false-positive and false-negative counts.

A speaker-level evaluation was also conducted to better evaluate the system's utility in clinical applications, where per-speaker diagnosis is more relevant than per-utterance classification. This was done using a majority voting scheme across each speaker's 33 utterances, where the majority class prediction determined the final speaker label (healthy or disordered). The results of this speaker-level evaluation are shown in Figure 3.7.

The speaker-level confusion matrices reveal that the SSL representations offer a significant advantage in producing consistent and reliable predictions across speakers. Notably, Wav2vec 2.0 Large achieved perfect speaker-level classification, correctly identifying all nine healthy and all eight disordered speakers. HuBERT and Wav2vec 2.0 Base also performed well, misclassifying only one or two speakers. In comparison, the standard representations exhibited weaker generalization. For example, openSMILE misclassified four healthy and three disordered speakers, while MFCC and eGeMAPS showed similarly high confusion, indicating their limitations.

These results suggest that SSL-based representations improve classification performance at the utterance level and offer significantly more robust speaker-level detection. This level of consistency and accuracy is critical for real-world deployment in clinical settings, where reliability across entire assessments is necessary for informed decision-making.

3.4.1 Conclusion and Future Work

In this work, I investigated the effectiveness of self-supervised learning (SSL) representations in the Automatic Speech Disorder Detection (ASDD) system for young children. The experiments were conducted by building ASDD systems with different representations, including three SSL representations (W2V2_Large, W2V2_Base, HuBERT) for predicting the utterance of healthy or speech sound disordered among young children. As a comparison,

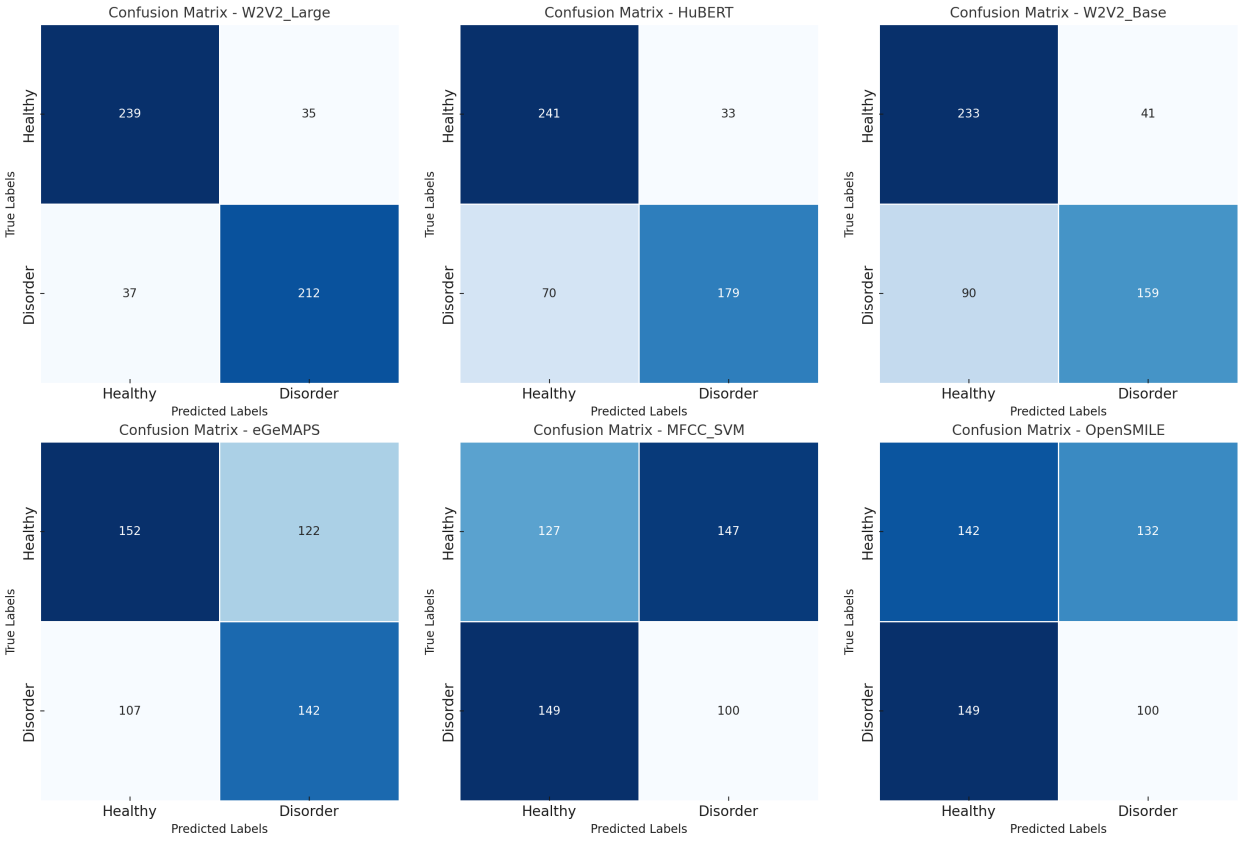

Figure 3.6: Confusion matrix on utterances

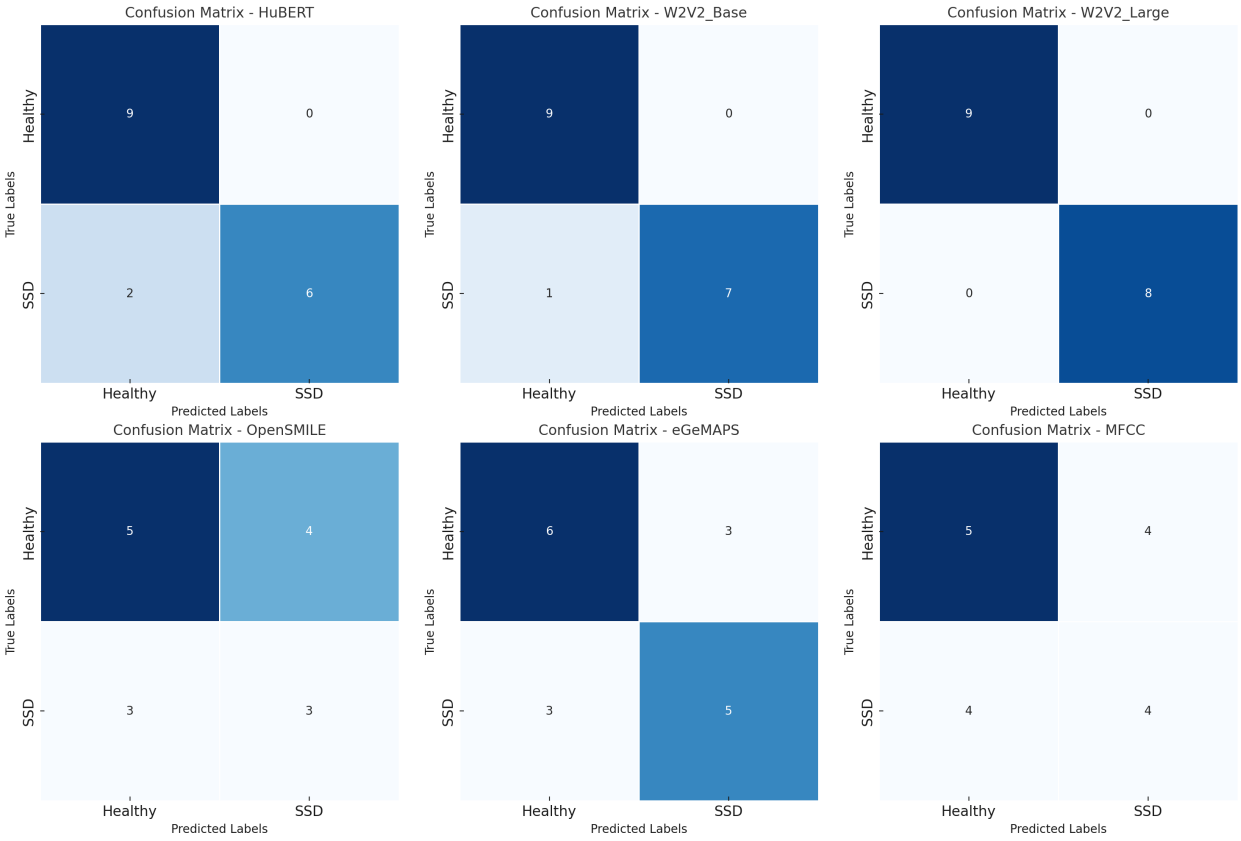

Figure 3.7: Confusion matrix of speaker level

I also developed ASDD systems with three standard representations (eGeMAPS, OpenSMILE, MFCC). The speech samples recorded from children aged 3–7 years old were used in this study.

The results from this study indicate that the SSL representations performed better than the standard representations in detecting children's speech sound disorder, with the W2V2_Large 25th representation achieving the best performance under five metrics. By comparing performances between layer-wise representations, it can be observed that later layers of SSL models may carry more information related to the speech sound disorder, although these models contain different architectures. This study also investigated the performance of an ASDD system in the context of low resources.

This study shows the application of SSL representations for young children in the ASDD system. In the future, this study can be extended to enhance the interpretability of SSL representations by integrating with explainable AI technologies or other features provided by speech-language pathologists. Furthermore, multi-classification for detecting more specific speech disorders beyond the binary classification of healthy and disordered is necessary to be investigated in the future.

Chapter 4

Study 3 - An assessment for the speech sound disorders with Self-supervised Learning (SSL) representation

4.1 Introduction

Quantifying pronunciation variation is a central challenge in automatic speech processing, with critical applications in second language (L2) assessment, dialect modeling, and clinical diagnosis of speech sound disorders (SSDs), especially in children [4, 43, 17]. Traditional approaches typically rely on symbolic representations, such as phoneme-level transcriptions, or engineered acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs)[4, 33]. However, these methods often fail to capture the rich, continuous, and highly variable phonetic information characteristic of spontaneous or impaired speech. Transcription-based approaches introduce subjectivity and are labor-intensive, while MFCCs offer limited sensitivity to prosodic and articulatory features that underlie human perception of speech variation.

Recent developments in self-supervised learning (SSL)[2, 48] have opened new avenues for extracting powerful, high-resolution speech representations from raw audio. Transformer-based models such as wav2vec 2.0 learn contextualized acoustic embeddings without relying on manual labels, offering a promising alternative to conventional handcrafted features. Prior studies [35, 48] have shown that such representations can effectively model fine-grained pronunciation differences and better correlate with human judgments of similarity and speech pronunciation compared to MFCCs and phoneme-edit distances. Notably, intermediate layers of these SSL models [7, 40], often encode generalizable acoustic-phonetic structures, making them suitable for tasks requiring sensitivity to pronunciation variation, such as accent similarity and speech disorder detection.

Furthermore, SSL embeddings have been successfully integrated into perceptual similarity spaces [8], in which speech trajectories are compared using techniques such as Dynamic Time Warping (DTW) to assess global speech pronunciation accuracy, particularly in L2 speakers. These approaches eliminate the need for phonetic segmentation or alignment and outperform MFCCs in modeling inter-speaker variation. Complementary work in clinical domains has also demonstrated the effectiveness of SSL embeddings [50] in automatically detecting SSDs via ASR and audio classification [43, 38, 16] pipelines, indicating their practical relevance in diagnostic and assistive technologies.

In this study, I investigated three approaches for developing an assessment tool for speech sound disorders using self-supervised learning (SSL) representations :

1. **SSL-based distance**, computed using layer-wise embeddings from wav2vec 2.0 and aligned using DTW;

2. **MFCC-based distance**, based on Euclidean or cosine distance between low-level acoustic feature vectors;

3. **Transcription-based distance**, calculated using edit distance between phoneme-level ground-truth or ASR-derived sequences.

I evaluate these distance measures in the context of children's speech, including those of typically developing children and those with clinically diagnosed SSDs. I aim to assess which method most accurately reflects perceptual differences in pronunciation and best supports downstream assessment of SSDs.

Key contributions of this work include:

- A systematic evaluation of three competing representations of speech variation (SSL, MFCC, transcription) within a unified framework;

- Empirical evidence demonstrating that SSL-based distances more closely reflect perceptual similarity, as reflected in superior separation of disordered vs. typical speech;

- Analysis of layer-wise representation quality in SSL models, supporting prior findings that intermediate layers yield the most informative embeddings for phonetic variation;
- Practical implications for designing interpretable, data-efficient pipelines for automated speech disorder screening.

This work bridges advances in SSL-based acoustic modeling with applications in speech health and education by grounding the evaluation in a real-world dataset of children's speech and focusing on clinically relevant distinctions. The findings show strong support for adopting SSL distance metrics in both research and deployment settings, and suggest promising directions for future work in model interpretability, cross-linguistic generalization, and low-resource assessment tools.

4.2 Related work

4.2.1 Literature Review

Neural representations for modeling variation in speech [4]

The authors comprehensively explored how neural-based acoustic representations, specifically those derived from self-supervised learning (SSL) models such as wav2vec 2.0, could effectively capture pronunciation variations. The authors investigated the capacity of these neural representations to quantify pronunciation differences across speakers, dialects, and languages, offering significant improvements over traditional methods such as phonetic transcriptions and MFCC-based acoustic measures.

This research's robust methodological framework extensively evaluated five self-supervised neural models: wav2vec, vq-wav2vec with BERT extension, wav2vec 2.0, XLSR-53, and DeCoAR. The authors systematically compared these neural models against traditional approaches, including phonetic transcription-based and MFCC-based methods, across multiple datasets containing both non-native American English speech and Norwegian dialectal variations. Their analysis convincingly demonstrated that transformer-based models, particularly

wav2vec 2.0, aligned more closely with human perceptual judgments of pronunciation similarity. For instance, wav2vec 2.0 significantly outperformed traditional approaches, achieving correlations up to 0.87 with human accent ratings, indicating superior efficacy in capturing nuanced pronunciation differences.

The paper also meticulously analyzed the influence of various experimental variables, such as the choice of reference speakers, speaker gender, and the linguistic background of speakers, further highlighting the robustness and generalizability of neural acoustic representations. Remarkably, the authors found that intermediate transformer layers often yielded better performance than final layers, suggesting that these layers encoded more generalizable acoustic properties rather than task-specific phoneme-level information optimized for speech recognition.

Additionally, the study presented a valuable visualization tool designed to elucidate where acoustic differences between pronunciations were most prominent. This visual interpretability aided in understanding the black-box nature of neural acoustic distance measures, providing insights into precisely which speech segments contributed most significantly to perceived pronunciation variation. Such tools offered intuitive feedback to users and held practical implications for applications in language learning, accent coaching, and clinical speech assessment.

However, the authors acknowledged certain limitations. They noted that transformer-based models trained primarily on English data demonstrated reduced effectiveness when applied to dialects or languages not extensively represented in the training datasets, as observed in the lower correlation scores for Norwegian dialect analysis. This finding pointed to the need for targeted pre-training or fine-tuning strategies to better align neural models with specific linguistic characteristics and dialectal contexts.

Overall, this paper significantly advanced the understanding of how SSL-based neural acoustic representations could effectively quantify pronunciation variations. It highlighted theoretical and practical implications, demonstrating the substantial advantages of using

transformer-based SSL models over conventional acoustic or phonetic transcription methods. The research set a clear foundation for future exploration into applying these robust acoustic representation techniques more broadly in speech pathology, dialectometry, and language teaching contexts.

A perceptual similarity space for speech based on self-supervised speech representations [8]

The authors of this paper introduce an innovative framework for quantifying speech similarity by leveraging self-supervised learning (SSL). They propose using a perceptual similarity space derived from SSL models to analyze acoustic-phonetic differences, particularly emphasizing second-language (L2) English speech pronunciation accuracy. This novel approach represents a significant shift from traditional acoustic analysis methods that depend on predefined acoustic parameters or manual transcription.

The paper's central contribution is the establishment of a perceptual similarity space created by employing SSL methods, specifically utilizing HuBERT representations. This methodology encodes speech utterances into multidimensional trajectories, which are then compared using dynamic time warping (DTW). This allows for meaningful distance measurement between utterances without requiring explicit phonetic segmentation or alignment to textual transcriptions. The authors effectively demonstrate that this approach accurately reflects perceptual judgments of speech pronunciation accuracy, particularly highlighting differences between native (L1) and non-native (L2) English speakers.

The experimental design is robust, involving speech recordings from multiple groups of L2 speakers (Mandarin, Korean, and Spanish heritage speakers), compared against L1 English speakers. The authors validate their perceptual similarity space method by showing a strong correlation between trajectory distances in the similarity space and speech pronunciation scores, as evaluated through word recognition accuracy by L1 English listeners. The

perceptual similarity distances significantly outperform traditional acoustic-phonetic measures such as mel-frequency cepstral coefficients (MFCCs) and vowel dispersion in predicting speech pronunciation accuracy.

Another strength is the paper's careful consideration of different SSL model layers, which reveals that intermediate layers often yield the most meaningful perceptual distinctions. This insight provides valuable guidance for future research and applications in SSL-based speech analysis. Moreover, the detailed exploration of inter-talker variability effectively demonstrates that trajectories for L2 speakers, due to varying proficiency levels, exhibit substantially more variability compared to those of L1 speakers. These findings highlight the power of the perceptual similarity space method in capturing nuanced and complex variations in speech production related to language proficiency.

However, the paper acknowledges some limitations. For instance, while the SSL-based perceptual similarity space excels in capturing global perceptual variations, it might overlook specific segment-level acoustic details better captured by traditional, hypothesis-driven acoustic analyses. The authors suggest future research directions, including integrating their approach with more targeted segment-level acoustic features or exploring the application of perceptual similarity spaces across other speech phenomena, such as speech impairments or different languages.

Overall, the study offers significant advancements in speech representation analysis, demonstrating the superior efficacy of SSL-derived perceptual similarity spaces in modeling acoustic variations relevant to human perception. This research advances theoretical understanding and suggests practical applications in educational, clinical, and technological settings, such as automated speech disorder detection and personalized language learning tools.

Automatic Detection of Speech Sound Disorder in Children Using Automatic Speech Recognition and Audio Classification [38]

In this paper, the authors address the challenging task of automatically identifying speech sound disorders (SSDs) in children using advanced deep learning methodologies. Specifically, they explore the integration of automatic speech recognition (ASR) and audio classification (AC) techniques, leveraging transformer-based models such as Whisper to assess pronunciation accuracy and diagnose SSDs with minimal human intervention.

The study was grounded in a substantial dataset collected from 573 children aged two to nine, recorded using standardized prompts from the Assessment of Phonology and Articulation for Children (APAC). Speech-language pathologists manually transcribed these recordings, identifying 92 children with SSDs based on the percent whole-word correct (PWC) metric. The authors used this meticulously annotated dataset to develop and rigorously evaluate five automated SSD detection methods—three based on ASR and two on AC models.

For the ASR-based approaches, the study carefully evaluated various strategies for determining diagnostic thresholds from automatic transcriptions. Specifically, the authors compared thresholds derived from human transcriptions, thresholds adjusted based on ASR accuracy, and a kernel density estimation (KDE) approach. Their findings showed that KDE-based thresholding achieved the most balanced performance by accounting for transcription inaccuracies, reaching an unweighted average recall (UAR) of 73.5%. Notably, the KDE method mitigated overestimation of SSD prevalence caused by systematic ASR errors, thereby improving diagnostic reliability.

The AC-based methods introduced an alternative diagnostic pathway. Two classification strategies were developed: speaker-level classification using combined samples, and word-level classification of individual target words. The word-level method outperformed all others, achieving the highest UAR of 73.9% and superior accuracy and F1 scores. This

success was attributed to the fine-tuned audio classification model, which attained 81.6% accuracy in identifying correctly pronounced words, substantially higher than the ASR model's 60.2%, and offered a robust alternative that reduced reliance on transcription.

The authors also provided detailed explanations of how the PWC metric was calculated. For ASR-based methods, PWC was derived from Whisper-generated transcriptions, adjusted to reflect real-world performance. For AC-based methods, PWC was based on the proportion of individual words classified correctly by the model. This methodological clarity enhances reproducibility and supports practical application of the proposed diagnostic tools.

Furthermore, the study addressed several limitations and offered specific directions for future work, including the need for larger datasets to strengthen age-based PWC distributions, fine-tuning ASR models on child speech data, and applying data augmentation to address class imbalance in AC tasks.

In summary, this paper advanced automated SSD detection by demonstrating the complementary strengths of ASR and audio classification methods. It offered a precise and reproducible framework for future research and highlighted the critical importance of robust model training and threshold calibration for clinical deployment.

4.3 Research Problems / Hypothesis

Research Questions:

- How can pronunciation variation in children's speech be effectively quantified for applications such as speech disorder detection? Children's speech is highly variable due to developmental differences in articulation, prosody, and speech motor control. Traditional distance measures, such as MFCC-based acoustic distances or phoneme-level transcription comparisons, may not adequately capture these nuances. There is a need to explore whether distances computed from modern self-supervised learning (SSL) embeddings can more accurately and robustly quantify pronunciation variation.

- How do SSL-based pronunciation distance measures compare with traditional methods (MFCC and transcription distance) in capturing speech variation relevant to speech sound disorders?

  While MFCC and transcription distances are well-established, they may not align with perceptual judgments or perform reliably on disordered or variable speech. This study investigates whether SSL-based distances offer better discrimination between typical and disordered speech patterns, particularly in children.

- Which layers of SSL models yield the most perceptually meaningful and diagnostically relevant speech distances?

  SSL models encode information across many layers, with different layers capturing different levels of acoustic or linguistic information. This research explores whether intermediate layers in models like wav2vec 2.0 provide more useful representations for capturing pronunciation differences than final task-tuned layers.

Proposed Solution:

In this study, I evaluate the variation in pronunciation using distances derived from three sources: self-supervised speech representations (SSL), cepstral coefficients of mel frequency (MFCCs), and phoneme-level transcription alignments. SSL embeddings are extracted from different transformer layers of models such as wav2vec 2.0. Pronunciation distance is computed using dynamic time warping (DTW) on SSL and MFCC features, and edit distance for phoneme transcriptions. These distances are then analyzed and compared for their ability to distinguish between children with and without speech sound disorders. Performance is evaluated regarding classification accuracy, robustness to variability, and alignment with perceptual intuitions.

Research Hypothesis:

I hypothesize that distance measures derived from self-supervised learning (SSL) representations will more accurately capture pronunciation variation in children's speech than

traditional approaches based on Mel-Frequency Cepstral Coefficients (MFCCs), which correlate more with phoneme-level transcriptions. Specifically, I expect:

- SSL-based distances will yield higher classification accuracy in distinguishing disordered vs. typical speech than MFCC or phoneme distances.
- Intermediate layers of SSL models will encode the most diagnostically relevant information to measure pronunciation similarity.
- SSL-based methods will be more robust to speaker identity, speaking rate, and word content variability.

4.4 Method

4.4.1 APTct for Computing Transcription Distance

To quantify pronunciation variation using symbolic representations, I incorporate the Automated Phonetic Transcription Comparison Tool (APTct) (Figure 4.1)to compute distances between phonetic transcriptions. APTct is a freely available, validated web-based application designed to compare phonetic transcriptions developed by [3]. It supports full IPA coverage via a graphical point-and-click keyboard and implements a robust phonologically informed version of the Levenshtein edit distance algorithm.

Modified Edit Distance Algorithm

The core of APTct is a modified edit distance framework that improves upon the standard Levenshtein algorithm by integrating critical linguistic principles when comparing phonetic transcriptions. The Levenshtein edit distance is a widely used algorithm for quantifying how different two sequences are by calculating the minimum number of single-character edits needed to transform one string into another. These edits can include insertions, deletions, or substitutions, each typically assigned a cost of 1. The algorithm operates by constructing a matrix where each cell represents the cumulative edit distance between prefixes of the

two strings. It uses dynamic programming to efficiently compute the optimal alignment by recursively comparing characters from both strings and choosing the minimum-cost operation at each step. For example, transforming the word “kitten” into “sitting” would involve three operations: substituting ’k’ with ’s’, substituting ’e’ with ’i’, and inserting a ’g’ at the end, resulting in a distance of 3. This method is helpful in speech processing tasks such as comparing phoneme transcriptions, evaluating speech recognition errors, or analyzing pronunciation differences. However, the basic Levenshtein distance treats all edits equally costly and does not consider linguistic context. Therefore, for applications like phonetic transcription comparison, more advanced versions of the algorithm, such as the one used in APTct, introduce weighted penalties and phonologically informed alignment rules to reflect perceptual and linguistic relevance better. The APTct tool extends the traditional Levenshtein edit distance algorithm by incorporating phonologically informed enhancements that improve its ability to capture meaningful pronunciation differences. First, it applies weighted penalties to distinguish between different types of edits: insertions, deletions, or substitutions of base phonemes (vowels or consonants) are assigned a cost of 1.0, while diacritic changes, which reflect more subtle phonetic features such as nasalization or devoicing, are given a reduced penalty of 0.5. This weighting reflects that diacritic differences represent more fine-grained articulatory variation and should not be penalized as heavily as phoneme substitutions. Secondly, APTct follows the nucleus alignment principle, which biases the alignment to prefer vowel-to-vowel and consonant-to-consonant matches. In contrast, vowel-consonant substitutions are considered less plausible and are penalized more heavily (with a cost of 2.0), ensuring that the computed distance reflects phonological structure more accurately. The algorithm also adheres to the strict order principle, which preserves the sequential order of phonemes during alignment, preventing unrealistic reordering of segments that might otherwise minimize distance at the cost of linguistic plausibility. Finally, APTct incorporates diacritic-aware processing, treating complex phonetic constructs such as diphthongs

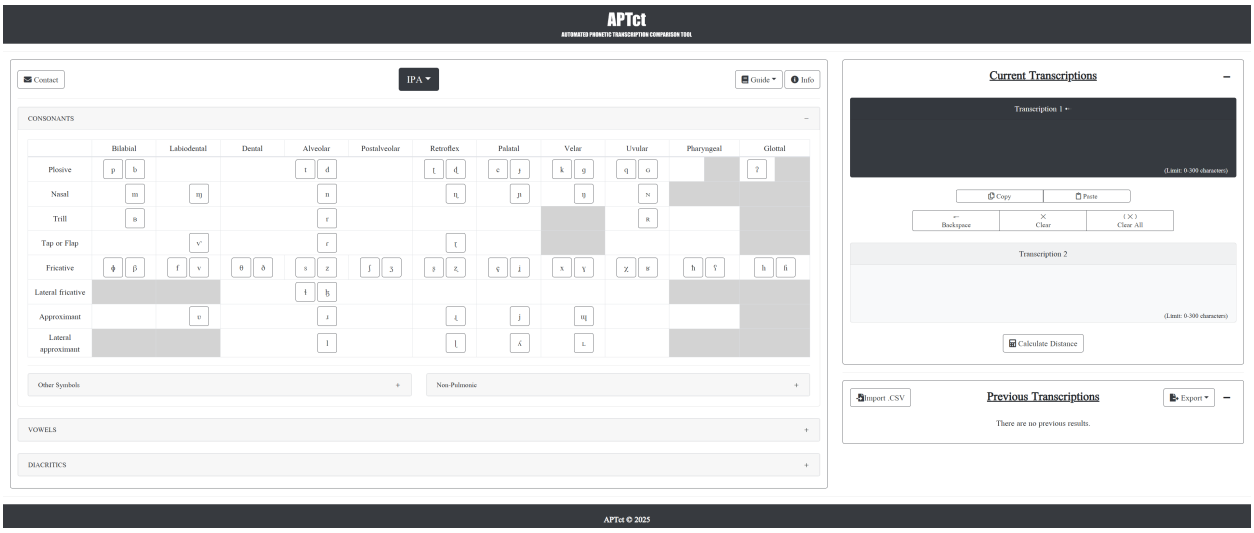

Figure 4.1: APTct

and affricates as unified phonological units through tie bars. It also handles syllabic consonants, suprasegmental features, and symbols from the extended IPA (extIPA), making it well-suited for analyzing disordered speech. These enhancements enable APTct to compute transcription distances that are quantitative, linguistically interpretable, and perceptually relevant. This experimental design uses APTct to compute transcription distances between each child's spoken production (manually transcribed in IPA) and the canonical target transcription. This process yields a quantitative measure of pronunciation deviation regarding phoneme-level and sub-phonemic changes. The computed distance reflects the number and type of changes needed to transform the child's production into the target form, accounting for substitutions, deletions, insertions, and diacritic differences.

APTct is publicly available at: **https://aptct.auburn.edu**

4.4.2 MFCC-Based Distance

Mel-frequency cepstral Coefficients (MFCCs) are among the most widely used features in speech signal processing, particularly in tasks involving speech recognition, speaker identification, and speech pathology assessment. MFCCs capture the short-term spectral envelope of a speech signal by mimicking the nonlinear frequency resolution of the human auditory

system. The core idea is to apply a filter bank spaced according to the mel scale, which emphasizes lower-frequency, more perceptually relevant components, followed by a Discrete Cosine Transform (DCT) to decorrelate the spectral features [5].

This study computed MFCCs for each audio frame, typically using a 25 ms window with a 10 ms hop. Each utterance results in a time series of MFCC feature vectors (e.g., 13-dimensional vectors over T frames). I compute the MFCC-based distance using Dynamic Time Warping (DTW) to quantify the pronunciation difference between two utterances. DTW is a temporal alignment technique that identifies the optimal (i.e., lowest-cost) non-linear alignment between two sequences of unequal length. It allows for local stretching or compression of time axes, which is especially useful for comparing speakers with different speaking rates or syllable durations.

The local distance between frames is typically computed using Euclidean distance or cosine similarity between MFCC vectors. The resulting DTW distance reflects the acoustic dissimilarity between two utterances, aggregated across all aligned frames. Smaller MFCC-DTW distances indicate more similar acoustic realizations.

However, while MFCCs effectively capture low-level spectral features such as formant structure and voicing characteristics, they are inherently limited in their ability to encode suprasegmental or prosodic information and do not explicitly model contextual dependencies or long-range phonetic patterns. Also, MFCCs are sensitive to speaker variability and noise, and may not align well with the human perception of pronunciation similarity, particularly in disordered or child speech. For these reasons, MFCC distance serves as a strong but limited baseline in this study, against which I compare more context-aware approaches such as SSL-based embeddings.

Dynamic Time Warping (DTW) is a time-series alignment algorithm that computes an optimal match between two sequences that may differ in length or local temporal structure. It is particularly well-suited for speech processing because spoken utterances often exhibit temporal variability — for example, two speakers may pronounce the same word at different

speeds, or with different prosodic patterns. DTW handles this variability by allowing non-linear alignments between time axes, essentially "warping" the sequences in time to find the lowest-cost alignment.

In pronunciation distance, each speech utterance is first converted into a sequence of acoustic feature vectors (e.g., MFCCs or SSL embeddings), one per short speech frame. DTW then computes a cost matrix, where each cell contains the distance (typically Euclidean or cosine) between a frame from one utterance and a frame from the other. A warping path is then computed through this matrix, starting from the first frame of both sequences and ending at the last, such that the total alignment cost is minimized. The DTW algorithm ensures that this path respects temporal continuity (no jumps or reversals) and monotonicity (progresses forward in time), making it suitable for aligning sequential acoustic content.

The result is a single scalar score representing the overall acoustic dissimilarity between two utterances, accounting for segmental differences (e.g., mispronunciations) and suprasegmental variation (e.g., timing and rhythm). The path length often normalizes this DTW score to control for utterance duration. In pronunciation research, DTW is commonly used to compare a speaker's production to a reference, either a canonical form or another speaker's output, making it a natural fit for applications such as automatic speech assessment, accent similarity, and speech sound disorder detection.

DTW is mighty when used with high-resolution representations like MFCCs or SSL embeddings, as it enables comparison without requiring phoneme-level segmentation or transcription. However, it assumes a frame-by-frame comparison and can be sensitive to irrelevant variations if the underlying features are not well normalized. In this study, DTW is applied to both MFCC and SSL-based sequences to compute pronunciation distances, enabling a direct comparison of how well each feature type supports perceptually relevant speech alignment.

4.4.3 Dataset

This study utilizes speech data from the Comprehensive Assessment of Articulation and Phonology (CAAP) module within the SEED (Speech Evaluation, Exploration, and Discovery) [45]corpus. The CAAP is a standardized clinical assessment tool that speech-language pathologists use to evaluate children's production of English consonants and phonological patterns. In the SEED corpus, the CAAP module includes high-quality audio recordings of children aged approximately 2.5 to 9 years, collected in structured elicitation sessions using picture-based prompts. Each child is asked to produce a fixed set of target words designed to cover a representative inventory of phonemes across various positions (initial, medial, and final) and phonotactic contexts.

A distinguishing feature of this dataset is the availability of manually transcribed phoneme-level annotations for each recorded word. These transcriptions, performed by trained linguists or speech-language pathologists using the International Phonetic Alphabet (IPA), capture canonical productions and typical phonological errors of speech sound disorders. Notably, the dataset includes diagnostic labels identifying whether each speaker has a speech sound disorder (SSD)or is typically developing (TD).

The data set contains multiple utterances of the exact target words in different speakers, allowing direct pairwise comparison of pronunciation. This makes it particularly well-suited for evaluating distance-based approaches to modeling pronunciation variation. This study uses the manually verified phoneme transcriptions to compute transcription-based distances via APTct. At the same time, the raw audio is processed to extract both MFCCs and self-supervised learning (SSL) embeddings for calculating acoustic-based distance measures. By leveraging the CAAP-in-SEED dataset's combination of raw audio, expert phonetic transcriptions, and clinical labels, I can systematically assess how well different distance metrics capture perceptually and diagnostically relevant variation in children's speech.

1. Duck
2. Shoe
3. Pig
4. Bed
5. Teeth
6. Dog
7. Cage
8. Gate
9. Mouse
10. Knife
11. King
12. Ring
13. House
14. Hive
15. Fish
16. Van
17. Seal
18. Zoo
19. Sheep
20. Jar
21. Cheese
22. Rake
23. Leaf
24. Watch
25. Web
26. Yoyo
27. Thumb
28. Bathe
29. Them
30. Clown
31. Flag
32. Glove
33. School
34. Snake
35. Swing
36. Bridge
37. Treasure
38. Computer
39. Dinosaur
40. Elephant
41. Grasshopper
42. Fingernail
43. Lemonade
44. Helicopter
45. Thermometer
46. Basketball

Figure 4.2: The Comprehensive Assessment of Articulation and Phonology (CAAP)

4.5 Implementation

I used the Automated Phonetic Transcription Comparison Tool (APTct) to compute transcription-based pronunciation distances, which provides a structured and linguistically informed method for aligning and evaluating phoneme-level differences between two TIMIT phoneme transcription sets [34]. As illustrated in Figure 4.3, APTct compares two phonetic strings—one representing the child's spoken output and the other the canonical target form. In this case, the child produced the sequence `/baesihbao/`, while the target transcription is `/baeskahtbaol/`. This subtle difference includes inserting an additional consonant `/t/` in the child's production.

The APTct tool performs a visual alignment, displaying both transcriptions horizontally and showing aligned segments in parallel rows. It highlights matching phonemes in green and marks mismatches using a color-coded scheme: deletions in red, insertions in blue, and substitutions in orange. In the technical alignment matrix, the alignment status of each segment is displayed along with its associated penalty score. In this example, all phonemes are perfectly aligned except for the extra /t/ in the child's transcription, which is not present in the target and is classified as a deletion, incurring a penalty of 1.0. All other aligned segments incur zero penalty, reflecting correct matches.

APTct's alignment algorithm follows an extended version of the Levenshtein edit distance that integrates phonological principles. It treats phoneme deletions, insertions, and substitutions with base penalties and accounts for linguistic detail. For example, diacritic differences (such as devoicing or nasalization) incur reduced penalties of 0.5, acknowledging their more subtle phonetic nature. Furthermore, nucleus alignment constraints penalize implausible substitutions, such as vowel-to-consonant mappings, more heavily (with a penalty of 2.0), while promoting vowel-vowel or consonant-consonant alignments.

The implementation of the MFCC-based pronunciation distance is performed within a function named `compute_speakers_distance`. This function takes as input two audio file dictionaries (`file1` and `file2`), FFT and hop parameters for feature extraction, the number

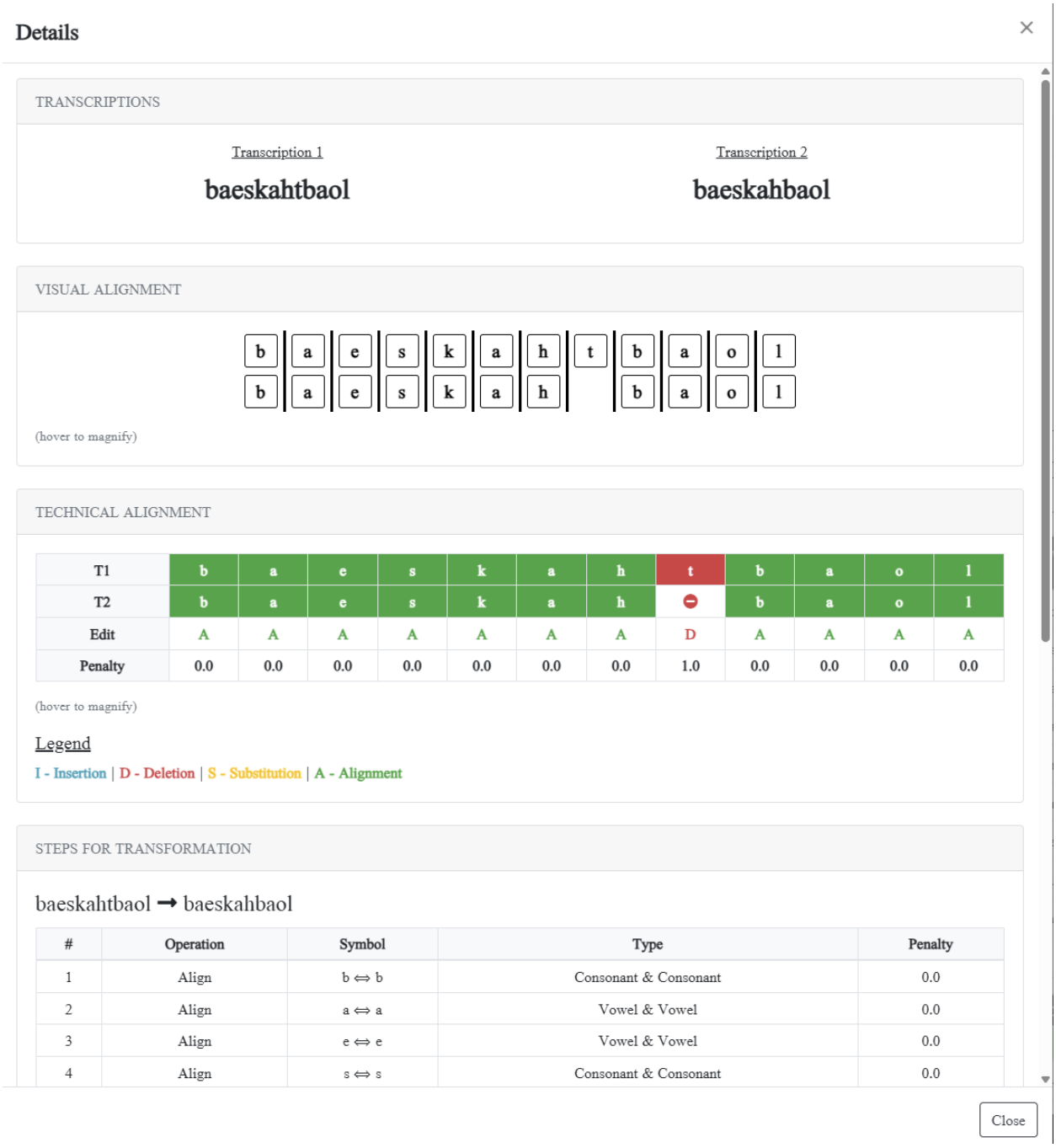

Figure 4.3: Phoneme distance by APTct

of MFCC bins, and a speaker group label. The function begins by extracting the raw audio waveforms and their corresponding sampling rates from the input dictionaries. It includes a safeguard to skip computation if either of the audio files is empty, returning a default value of -1 in such cases.

Next, MFCC features are extracted for each audio signal using the `librosa.feature.mfcc()` function. The function parameters include the raw audio waveform (`y`), the sampling rate (`sr`), FFT window size (`n_fft`), hop size (`hop_length`), and the number of MFCC coefficients (`n_mfcc`, defined by `n_bins`). After extraction, cepstral mean and variance normalization (CMVN) is applied independently to each MFCC matrix by subtracting the mean and dividing by the standard deviation. This normalization helps reduce variability due to speaker identity, microphone differences, or recording conditions.

After preprocessing, the function calculates the acoustic distance between the two MFCC sequences using Dynamic Time Warping (DTW), implemented via the `lib_dtw` function. DTW computes an optimal non-linear alignment path between the two feature matrices and returns the accumulated cost matrix. The function selects the total cost of the optimal warping path (i.e., the last cell in the DTW matrix, `[-1,-1]`) as the final distance. Additionally, there is a conditional clause: if the current sample belongs to group `'BT'` (potentially "between-group") and the speaker has a speech disorder (marked by `'DisorderOrNot' == 'Y'`), the distance is computed using `lib_dtw(mfcc2, mfcc1)` to maintain consistent reference directionality. Otherwise, it defaults to `lib_dtw(mfcc1, mfcc2)`.

This implementation allows for robust pairwise comparison of speech recordings, capturing both segmental and suprasegmental acoustic variation. The resulting DTW-based MFCC distance is a baseline for quantifying pronunciation differences and is used in further analysis and classification tasks within the study.

Table 4.1: Pearson Correlation coefficients

Method	Pearson Correlations
MFCC	0.301
W2V2(18)	0.75

4.6 Results

Table 4.1 compares the Pearson correlation coefficients between model-derived embedding distances and phonetic distances, offering insight into how well two speech feature extraction methods—MFCC and Wav2Vec2—preserve phonetic information.

The MFCC (Mel-Frequency Cepstral Coefficients) method correlates 0.301 with phonetic distances. MFCCs are handcrafted features that model the human auditory system's sensitivity to different frequency bands. While they are widely used in traditional speech processing tasks, their relatively low correlation indicates that MFCCs only partially capture phonetic similarity. This is likely because MFCCs primarily reflect spectral envelope characteristics without incorporating contextual or higher-level linguistic information, which is critical for capturing subtle phonetic differences, especially in disordered or accented speech.

In contrast, Wav2Vec2 embeddings, specifically from layer 23, show a substantially higher correlation of 0.75 with phonetic distances. This strong correlation indicates that Wav2Vec2 captures rich and nuanced phonetic information in its intermediate representations. Wav2Vec2 is a self-supervised learning model trained on large amounts of unlabeled audio data to predict masked speech segments. Its deep architecture allows it to model long-range temporal dependencies and contextual information, essential for accurately reflecting phonetic structure. The selection of layer 18 likely corresponds to an optimal point in the model where phonetic-level abstractions are maximally represented before the representations become more task-specific in higher layers.

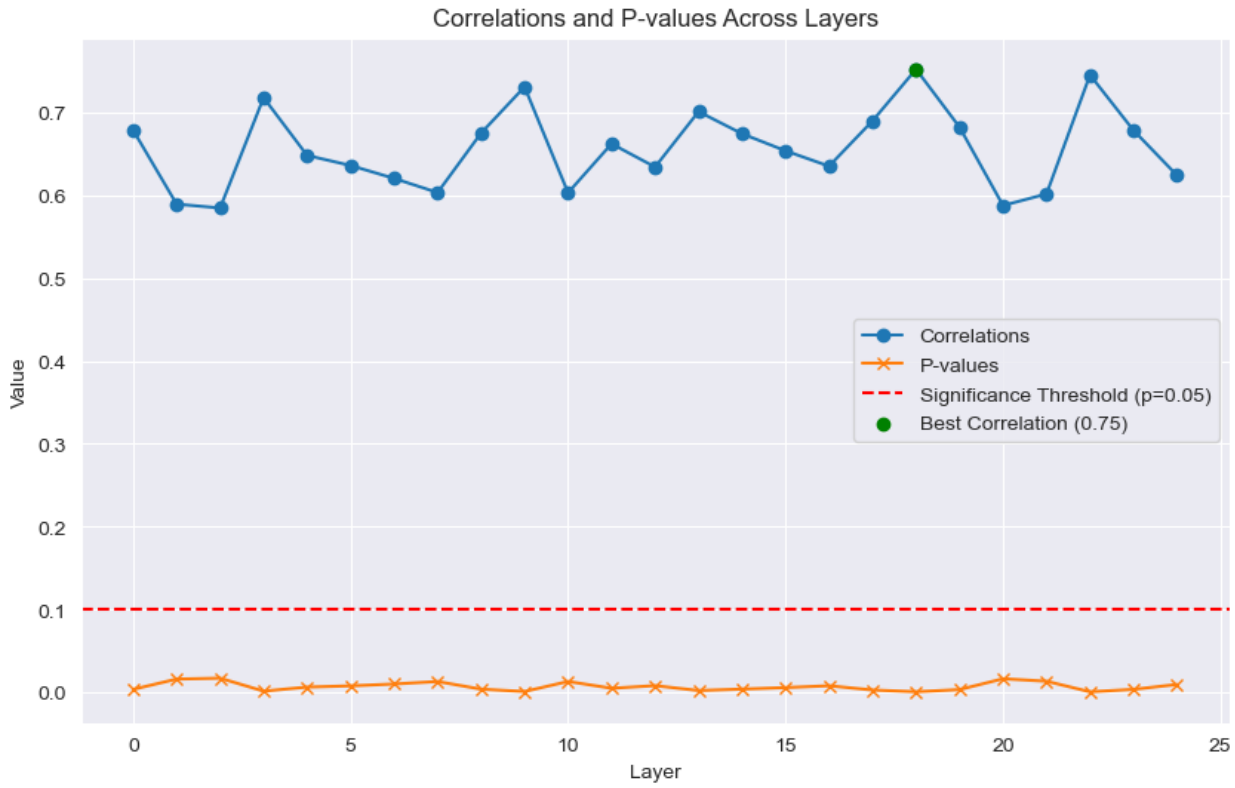

Figure 4.4: Correlations and P-values Across Layers

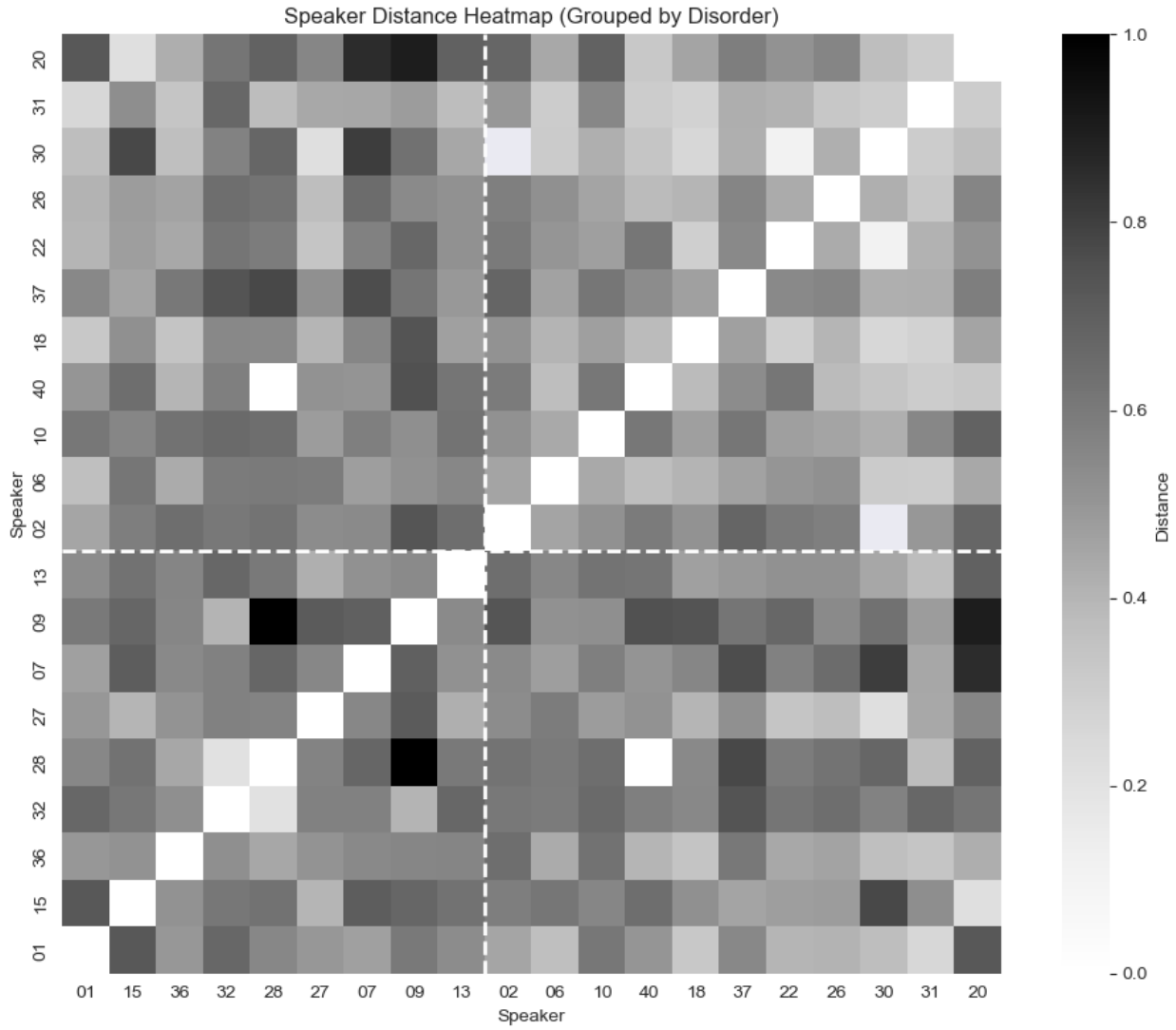

Figure 4.5: Speaker Distance Heatmap

Figure 4.4 illustrates the correlation coefficients and corresponding p-values between embedding distances derived from each layer of the Wav2Vec2 model and phonetic transcription-based distances. Correlation values fluctuate moderately across the 25 layers analyzed, ranging from approximately 0.55 to 0.75. Notably, layer 18 yields the strongest correlation (0.75), distinctly marked by a green indicator, signifying the layer most closely aligned with phonetic distances. Meanwhile, p-values across all layers are consistently very low, approaching zero and remaining substantially below the standard significance threshold ($p = 0.05$), highlighting the statistical reliability of these correlations. Overall, this analysis demonstrates that the representations extracted from all layers of the Wav2Vec2 model significantly reflect phonetic similarity, with layer 18 providing the most robust alignment.

Figure 4.5 presents a visualization of speaker distances calculated using embeddings derived from the Wav2Vec2 model. Speakers are organized into two groups: disordered speakers positioned in the bottom-left quadrant and non-disordered speakers in the top-right quadrant, separated by dashed white lines. The color intensity reflects embedding distances, with lighter shades indicating larger distances (greater dissimilarity) and darker shades representing smaller distances (higher similarity). As expected, a clear diagonal pattern emerges in both groups, indicating minimal distances for self-comparisons. Within-group distances, defined by regions closer to the diagonal within each quadrant, generally show darker shades, indicating greater similarity among speakers of the same group. In contrast, distances between groups (off-diagonal quadrants) display lighter shades, suggesting pronounced differences in embeddings between disordered and non-disordered speakers. Overall, this heatmap visually confirms the ability of embeddings to effectively distinguish between disordered and non-disordered speakers based on acoustic-phonetic characteristics captured by the Wav2Vec2 model.

Figure 4.6 illustrates the distribution of t-SNE distances between three pairs of speakers: N-N, Y-Y and Y-Y using a box plot. Speaker pairs are grouped as N-N (both normal), N-Y

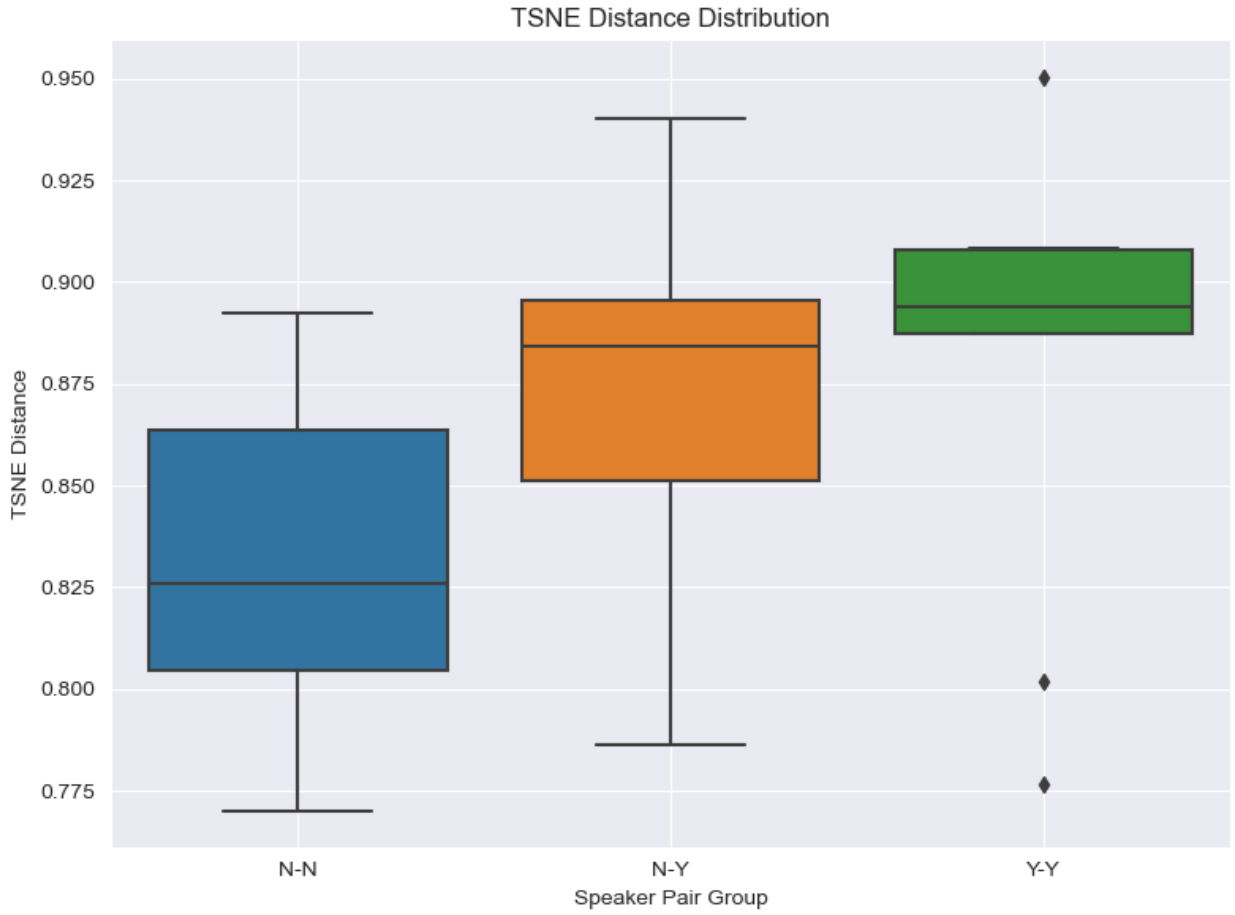

Figure 4.6: T-SNE Distance Distribution: speaker pairs are grouped as N-N (normal-normal), N-Y (normal-disordered), and Y-Y (disordered-disordered) to compare t-SNE distance distributions and assess separability in the embedding space.

(one standard, one disordered), and Y-Y (both disordered) to assess how speech representations differ across these categories. The t-SNE distance quantifies the similarity between speaker embeddings in a reduced two-dimensional feature space, where lower distances suggest more similar speech representations and higher values indicate greater dissimilarity.

The N-N group (blue box) consists of pairs in which both speakers are from the normal population. The median t-SNE distance in this group is approximately 0.83, and the interquartile range (IQR) spans roughly from 0.80 to 0.86. The whiskers extend from around 0.77 to 0.89, with no extreme outliers visible. This distribution shows that, while there is some variability, regular speakers tend to cluster relatively closely in the t-SNE space, suggesting consistency in their speech representations.

The N-Y group (orange box) includes mixed pairs with one normal and one disordered speaker. This group exhibits a slightly higher median t-SNE distance, around 0.88, indicating increased dissimilarity compared to the N-N group. The IQR for this group ranges approximately from 0.85 to 0.90, and the whiskers extend from 0.78 to 0.94. The wider range and higher median suggest introducing a disordered speaker leads to a more distinct separation in the feature space. However, this group also shows moderate variability, implying that dissimilarity varies across mixed pairs.

The Y-Y group (green box) includes pairs of disordered speakers. It has the highest median distance, around 0.89, and the narrowest IQR, roughly 0.88 to 0.91, indicating a tightly clustered distribution. Although a few outliers appear, one as low as 0.775 and another as high as 0.95, the main body of data is highly consistent. This suggests that disordered speakers form a distinct and compact cluster in the embedding space, characterized by a uniform deviation from standard speech patterns.

$$\text{Pronunciation Score} = Normalized(1 - \frac{D_{APTct}}{\max(|T_1|, |T_2|)}) \tag{4.1}$$

In this 4.1, the Levenshtein distance is computed using the APTct web-based interface, a specialized tool for evaluating phonetic transcription similarity. APTct takes two phonetic

transcriptions—typically a target (reference) and a hypothesis (produced by a speaker)—and calculates the Levenshtein distance, which quantifies the minimum number of edit operations (insertions, deletions, or substitutions) needed to convert one string into the other. Unlike orthographic comparison tools, APTct operates at the phoneme level, using the Alphabet transcription to provide a linguistically meaningful measurement of transcription accuracy.

This raw distance is then normalized by dividing it by the length of the longer transcription, allowing for comparisons across transcriptions of varying lengths. Such normalization ensures that longer words or utterances do not inherently yield higher distances, thus enabling fair speech pronunciation assessments across different samples. The resulting normalized score reflects the relative phonetic dissimilarity between two transcriptions, which is then transformed into a scaled speech pronunciation score ranging from 0 to 5. In this scale, a score of 5 indicates a perfect match (i.e., zero edit distance). In contrast, lower scores reflect increasing degrees of phonetic deviation, thereby serving as a proxy for perceived speech pronunciation.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

y_i = true value , $\hat{y}_i$ = predicted value, n = number of data points

To evaluate the effectiveness of SSL representations in predicting speech pronunciation score, I employed linear regression models using distance features derived from MFCCs and Wav2vec2 embeddings. These distance features served as predictors of the speech pronunciation score, which is scaled by the transcription distance. Model performance was assessed using the Mean Squared Error (MSE), which quantifies the average squared difference between predicted and actual speech pronunciation scores. Lower MSE values indicate more accurate predictions, reflecting a stronger correspondence between acoustic characteristics and perceived speech pronunciation accuracy.

Figure 4.7 compares the distributions of Mean Squared Error (MSE) for speech pronunciation score prediction using MFCC-based and Wav2Vec2-based distance features within a

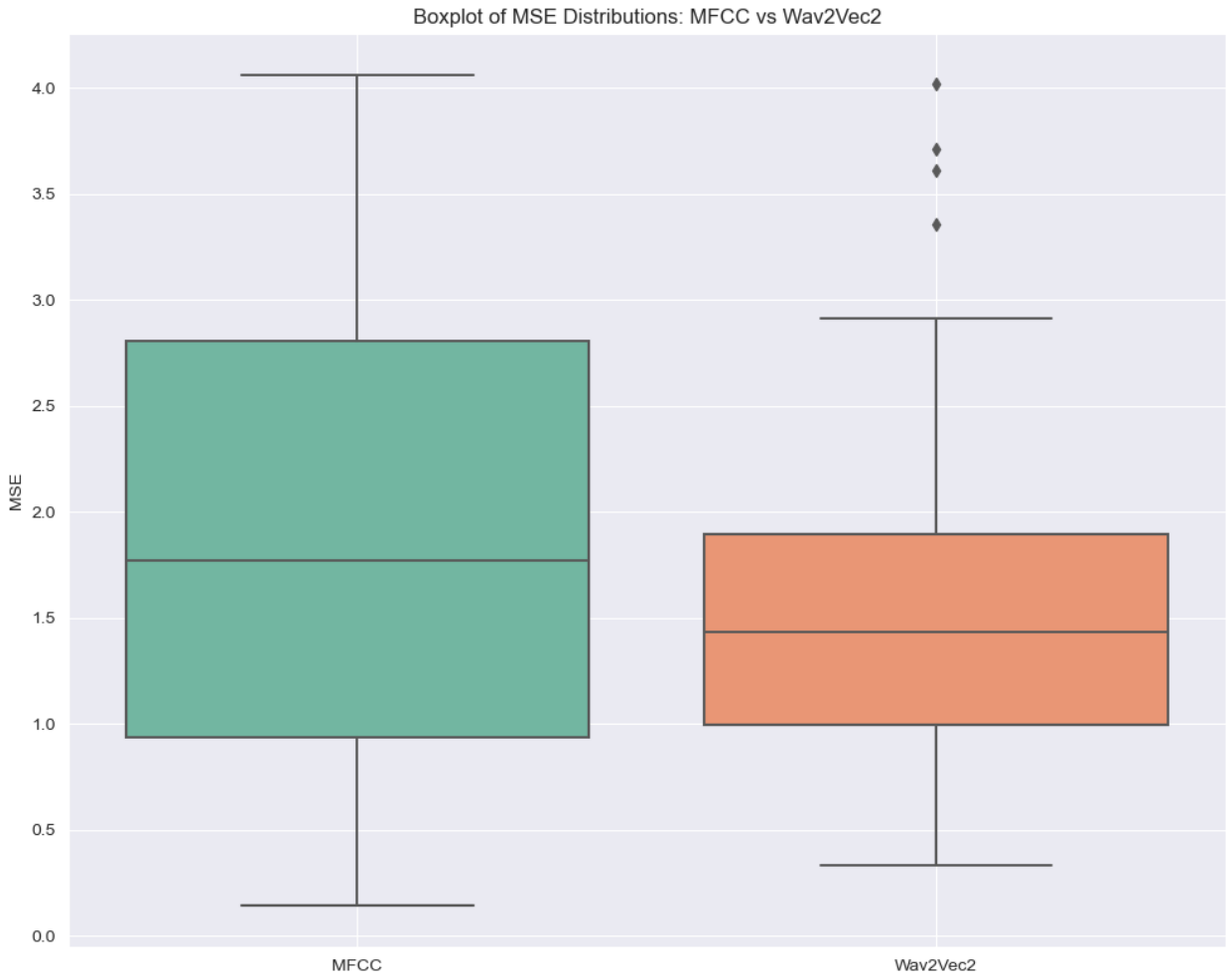

Figure 4.7: Box-plot of MSE Distributions

linear regression framework. The Wav2vec2 model exhibits a noticeably lower median MSE and a narrower inter-quartile range (IQR) than the MFCC model, indicating more accurate and consistent predictions across word items. Notably, several outliers are present in the Wav2vec2 distribution. These correspond to specific words for which the model's predictions deviated substantially from the human-rated speech pronunciation scores. When applied to disordered or out-of-distribution speech, such outliers may result from phonetic complexity, atypical articulation, or representational limitations of the Wav2vec2 embeddings. Despite these cases, the overall spread of error in Wav2vec2 is substantially smaller than that of MFCC, suggesting that self-supervised features provide a more robust basis for modeling perceived speech pronunciation.

4.7 Conclusion

This study compared MFCC and Wav2vec2-based speech representations across several dimensions—phonetic alignment, speaker separability, and speech pronunciation prediction accuracy. The results consistently demonstrate the superiority of Wav2vec2 embeddings in capturing linguistically meaningful and perceptually relevant information from speech.

First, correlation analyses between model-derived distances and phonetic transcription distances reveal that Wav2vec2 embeddings substantially outperform MFCCs in preserving phonetic structure. While MFCCs achieved a relatively low Pearson correlation ($r = 0.301$), Wav2vec2 embeddings from intermediate layers, particularly layer 18, reached correlations as high as $r = 0.75$. This suggests that the intermediate representations in Wav2vec2 align closely with phonetic-level abstractions, likely due to the model's ability to integrate long-range temporal and contextual information during pretraining. In contrast, MFCCs, as handcrafted features, primarily capture static spectral envelope properties and lack the expressiveness needed to represent complex phonetic distinctions, especially in disordered or accented speech.

The speaker distance heatmap further supports the phonetic sensitivity of Wav2Vec2. Embeddings show substantial intra-group similarity and clear inter-group separation between disordered and non-disordered speakers. This spatial organization implies that the learned representations capture acoustic-phonetic patterns systematically altered in disordered speech. Additionally, the t-SNE distance distribution reveals graded separability across speaker pair types (N-N, N-Y, Y-Y), where increasing phonetic divergence corresponds to greater distances in the embedding space. This further highlights the sensitivity of Wav2Vec2 to clinically relevant speech variability.

Regarding speech pronunciation prediction, linear regression models using Wav2vec2 distances produced significantly lower and more stable MSEs than those using MFCCs. While MFCC-based predictions exhibited greater variability and higher median error, Wav2vec2 results were more accurate and consistent across word types. Although a few outliers were observed in the Wav2vec2 boxplot, likely due to disordered articulations or out-of-distribution acoustic patterns, the overall distribution remained tightly centered. This indicates that Wav2vec2 not only preserves phonetic structure better but also generalizes well to perceptual tasks such as speech pronunciation scoring.

Collectively, these findings highlight the strengths of self-supervised learning in speech representation. Wav2vec2 embeddings provide a rich, context-sensitive basis for modeling linguistic similarity and clinical variation in speech. The results suggest that future work will explore fine-tuning strategies and broader clinical datasets to enhance model generalizability and interpretability further.

Chapter 5

Conclusion

This dissertation tackles significant challenges in automatic speech disorder detection (ASDD) for children by creating and assessing deep learning frameworks that utilize modern representation learning methodologies. Traditional ASDD systems, primarily based on handcrafted features like Mel-Frequency Cepstral Coefficients (MFCCs), frequently fail to capture the complex and context-dependent nature of evolving speech, particularly within children's disordered speech. To address these challenges, this research systematically explores the application of transformer-based architectures (ViT) and self-supervised learning (SSL) models, providing a cohesive view on enhancing the precision, interpretability, and clinical relevance of ASDD systems.

The first study uses the Vision Transformer (ViT) to classify disordered versus non-disordered speech using MFCC feature maps. By conceptualizing MFCCs as two-dimensional time-frequency "images," this method leverages ViT's attention-based patch representation to capture local and global acoustic dependencies. The ViT model surpasses traditional approaches such as SVMs and CNNs, especially in speaker-independent scenarios. This performance showcases that attention mechanisms can derive meaningful representations even from fixed, low-level features, linking traditional ASDD workflows with advanced deep learning models.

The second study demonstrates that self-supervised learning (SSL) models such as wav2vec 2.0 and HuBERT provide a substantial performance advantage over traditional handcrafted acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), eGeMAPS, and openSMILE, in the task of speech disorder detection. A detailed layer-wise analysis

identified the specific layer that effectively captures relevant characteristics essential for distinguishing disordered speech. Classifiers leveraging these SSL-derived representations consistently achieved higher accuracy and improved generalization across speakers, emphasizing the limitations of conventional features in modeling the complexity of disordered speech in children. These findings reinforce the potential of SSL models as powerful and generalizable tools for clinical speech assessment, particularly in low-resource and pediatric settings.

The third study advocates a transcription-free, perceptual similarity approach for evaluating speech disorder severity and pronunciation accuracy. Using dynamic time warping (DTW) on high-dimensional SSL embeddings, the framework measures pairwise distances between speech samples without text or phonetic transcriptions. These measurements correlate strongly with phonetic transcription-based Levenshtein distances and speech pronunciation accuracy. Layer-wise correlation analysis indicates that intermediate SSL layers (e.g., layer 18 of wav2vec 2.0) align best with phonetic similarity. Additionally, t-SNE visualizations and speaker-level heatmaps illustrate clear separations between disordered and non-disordered speakers, affirming the embeddings' capability to capture clinically significant acoustic-phonetic variations. Linear regression models leveraging SSL-based distance features achieve lower mean squared errors (MSE) in predicting speech pronunciation accuracy compared to MFCC models, confirming the robustness of SSL features for modeling perceptual speech characteristics.

These studies collectively comprehensively evaluate transformer-based and SSL approaches in pediatric ASDD. By integrating classification accuracy, interpretability, and perceptual similarity assessments, a holistic framework for detecting and assessing speech disorders is established. Demonstrating that SSL models surpass conventional methods across multiple facets and provide interpretable, transferrable results, this dissertation lays a technical and conceptual groundwork for future research and clinical application. Proposed methods could be integrated into early detection systems, progress monitoring tools, or

mobile health applications, offering scalable access to speech assessment for children. Looking ahead, future work might tailor pretrained SSL models to pediatric or clinical speech datasets to enhance responsiveness to developmental differences. Incorporating multi-modal data (e.g., articulatory kinematics, prosody, or facial gestures) could further boost performance and interpretability. This dissertation's tools and insights ultimately bridge machine learning progress with practical clinical impact in speech and language pathology.

Bibliography

[1] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun. An enhanced speech emotion recognition using vision transformer. *Scientific Reports*, 14(1):13126, 2024.

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

[3] Dallin J Bailey, Marisha Speights Atkins, Ishaan Mishra, Sicheng Li, Yaoxuan Luan, and Cheryl Seals. An automated tool for comparing phonetic transcriptions. *Clinical Linguistics Phonetics*, 36(6):495–514, 2022.

[4] Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137, 2022.

[5] Martijn Bartelds, Caitlin Richter, Mark Liberman, and Martijn Wieling. A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence*, 3:39, 2020.

[6] Eric Carle Bill Martin. *Brown bear, brown bear.* 1984.

[7] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. *ArXiv*, abs/1604.00825, 2016.

[8] Bronya R Chernyak, Ann R Bradlow, Joseph Keshet, and Matthew Goldrick. A perceptual similarity space for speech based on self-supervised speech representations. *The Journal of the Acoustical Society of America*, 155(6):3915–3929, 2024.

[9] François Chollet. Keras: The python deep learning library , type = conference proceedings.

[10] Philippe H. Dejonckere, Patrick J. Bradley, Pais Clemente, Guy Cornut, Lise Crevier-Buchman, Gerhard Friedrich, Paul van de Heyning, Marc Remacle, and Virginie Woisard. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology*, 258:77–82, 2001.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[13] Ö Eskidere and A. Gürhanlı. Voice disorder classification based on multitaper mel frequency cepstral coefficients features. *Comput Math Methods Med*, 2015:956249, 2015. 1748-6718 Eskidere, Ömer Gürhanlı, Ahmet Evaluation Study Journal Article United States 2015/12/19 Comput Math Methods Med. 2015;2015:956249. doi: 10.1155/2015/956249. Epub 2015 Nov 22.

[14] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, and Shrikanth S. Narayanan. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202 , ISSN = 1949–3045, 2015.

[15] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor , publisher = association for computing machinery, 2010.

[16] Javanmardi Farhad, Kadiri Sudarsana Reddy, and Alku Paavo. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, 158:103047, 2024.

[17] Yaroslav Getman, Ragheb Al-Ghezi, Ekaterina Voskoboinik, Tamás Grósz, Mikko Kurimo, Giampiero Salvi, Torbjørn Svendsen, and Sofia Strömbergsson. Wav2vec2-based speech rating system for children with speech sound disorder , booktitle = interspeech , publisher = international speech communication association (isca).

[18] Yuan Gong, Yu-An Chung, and James R. Glass. Ast: Audio spectrogram transformer. *ArXiv*, abs/2104.01778, 2021.

[19] Yuan Gong, Yu-An Chung, and James R. Glass. Ast: Audio spectrogram transformer. *ArXiv*, abs/2104.01778, 2021.

[20] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence , ISSN = 0162-8828*, 2024.

[21] Agustinus Bimo Gumelar, Eko Mulyanto Yuniarno, Wiwik Anggraeni, Indar Sugiarto, Vincentius Raki Mahindara, and Mauridhi Hery Purnomo. Enhancing detection of pathological voice disorder based on deep vgg-16 cnn , booktitle = 2020 3rd international conference on biomedical engineering (ibiomed) , publisher = ieee.

[22] Wei Han, Cheong-fat Chan, Oliver Chiu-sing Choy, and Kong Pang Pun. An efficient mfcc extraction method in speech recognition. *2006 IEEE International Symposium on Circuits and Systems*, pages 4 pp.–, 2006.

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460 , ISSN = 2329–9290, 2021.

[24] Ray D. Kent. Hearing and believing. *American Journal of Speech-Language Pathology*, 5(3):7–23 , DOI = doi 10.1044/1058–0360.0503.07, 1996.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[26] H. B. Klein and M. Liu-Shea. Between-word simplification patterns in the continuous speech of children with speech sound disorders. *Lang Speech Hear Serv Sch*, 40(1):17–30 , ISSN = 0161–1461 (Print) 0161–1461 , DOI = 10.1044/0161–1461(2008/08–0008), 2009.

[27] Vicsi Klára, Imre Viktor, and Mészáros Krisztina. Voice disorder detection on the basis of continuous speech , series = 5th european conference of the international federation for medical and biological engineering , publisher = springer berlin heidelberg.

[28] Alkis Koudounas, Gabriele Ciravegna, Marco Fantini, Giovanni Succo, Erika Crosetti, Tania Cerquitelli, and Elena Baralis. Voice disorder analysis: a transformer-based approach. *arXiv preprint arXiv:2406.14693*, 2024.

[29] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series.

[30] Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. Analysis of self-supervised speech models on children's speech and infant vocalizations. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 550–554. IEEE, 2024.

[31] Jueting Liu, Chang Ren, Yaoxuan Luan, Sicheng Li, Tianshi Xie, Cheryl Seals, and Marisha Speights Atkins. Speech disorders classification by cnn in phonetic e-learning system. Artificial Intelligence in HCI, pages 557–566. Springer International Publishing.

[32] Jueting Liu, Chang Ren, Yaoxuan Luan, Sicheng Li, Tianshi Xie, Cheryl Seals, and Marisha Speights Atkins. Speech disorders classification by cnn in phonetic e-learning system , series = artificial intelligence in hci , publisher = springer international publishing.

[33] Jueting Liu, Marisha Speights, Dallin J Bailey, Sicheng Li, Huanyi Zhou, Yaoxuan Luan, Tianshi Xie, and Cheryl D. Seals. Speech disorders classification in phonetic exams with mfcc and dtw, 2021.

[34] Carla Lopes and Fernando Perdigao. Phone recognition on the timit database. *Speech Technologies/Book*, 1:285–302, 2011.

[35] M. Mahendran, R. Visalakshi, and S. Balaji. Dysarthria detection using convolution neural network. *Measurement: Sensors*, 30:100913, 2023.

[36] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python , booktitle = scipy , type = conference proceedings.

[37] Juan M. Montero and Juan Carlos Martınez-Castrillo. Prediction of the degree of parkinson's condition using recordings of patients' voices , booktitle = proceedings of the ninth international conference on soft computing and pattern recognition (socpar 2017) , publisher = springer.

[38] Si-Ioi Ng, Cymie Wing-Yee Ng, Jiarui Wang, and Tan Lee. Automatic detection of speech sound disorder in child speech using posterior-based speaker representations. *arXiv preprint arXiv:2203.15405*, 2022.

[39] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model , booktitle = 2021 ieee automatic speech recognition and understanding workshop (asru) , publisher = ieee.

[40] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models , booktitle = icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp) , publisher = ieee.

[41] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830 , ISSN = 1532–4435, 2011.

[42] Wayne Secord and JoAnn S. Donohue. *CAAP-2: Clinical assessment of articulation and phonology-2 , publisher = Super Duper Publications , ISBN = 1607231220.* 2014.

[43] Mostafa Shahin, Usman Zafar, and Beena Ahmed. The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):400–412 , ISSN = 1932–4553, 2019.

[44] Irum Sindhu and Mohd Shamrie Sainin. Automatic speech and voice disorder detection using deep learning—a systematic literature review. *IEEE Access*, 12:49667–49681 , ISSN = 2169–3536 , DOI = 10.1109/access.2024.3371713, 2024.

[45] Marisha Speights Atkins, Dallin J. Bailey, and Suzanne Boyce. Speech exemplar and evaluation database (seed) for clinical training in articulatory phonetics and speech science. *Clinical Linguistics & Phonetics*, 34:878 – 886, 2020.

[46] Mohammad Tami, Sari Masri, Ahmad Hasasneh, and Chakib Tadj. Transformer-based approach to pathology diagnosis using audio spectrogram. *Information*, 15(5 , ISSN = 2078-2489 , DOI = 10.3390/info15050253), 2024.

[47] S. Tirronen, S. R. Kadiri, and P. Alku. The effect of the mfcc frame length in automatic voice pathology detection. *J Voice , ISSN = 1873-4588 (Electronic) 0892-1997 (Linking) , DOI = 10.1016/j.jvoice.2022.03.021*, 2022.

[48] Saska Tirronen, Farhad Javanmardi, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. Utilizing wav2vec in database-independent voice disorder detection , booktitle = icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp) , publisher = ieee.

[49] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*.

[50] Daniela A. Wiepert, Rene L. Utianski, Joseph R. Duffy, John L. Stricker, Leland R. Barnard, David T. Jones, and Hugo Botha. Speech foundation models in healthcare: Effect of layer selection on pathological speech feature prediction , type = journal article.

[51] Elisabeth H. Wiig, Wayne A. Secord, and Eleanor Semel. *CELF-Preschool-2: Clinical evaluation of language fundamentals, preschool , publisher = Harcourt Assessment , ISBN = 0158035356*. 2004.

[52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[53] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, and Guan-Ting Lin. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.