Advances in Consumer Research
https://acr-journal.com/

Volume-2 | Issue-5 | November 2025

Original Researcher Article

Auto DEAP: CNN-Transformer based hybrid model for automated pediatric speech misarticulation detection

Raghav Gohil[1*], Jinnal Raghwani[2], Rasika Adiseshan[3], Kanak Kadulkar[4], Gargi Rane[5], Yashvi Savla[6], Dr. Aruna U. Gawade[7], Dr Nilesh T. Rathod[8]

[1*]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 88793 67076, raghavgohil2004@gmail.com
[2]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 93214 61916, jinnal1811@gmail.com
[3]Department of Artificial Intelligence and Machine Learning, Dwarkadas.J.Sanghvi College of Engineering, Mumbai, India, +91 703028700, rasika.190804@gmail.com
[4]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 8530126999, kanakkadulkar@gmail.com
[5]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 82750 67935, ranegargi18@gmail.com
[6]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 9321215477, savlayashvid@gmail.com
[7]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 98190 03578, aruna.gawade@djsce.ac.in
[8]Department of Artificial Intelligence and Machine Learning, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India, +91 99870 63437, nilesh.rathod@djsce.ac.in

Received: 02/10/2025
Revised: 31/10/2025
Accepted: 08/11/2025
Published: 13/11/2025

ABSTRACT

Speech is an integral part of a child's growing years; it is how a child learns to express. Speech misarticulation disorders, when left unattended, can cause problems in a child's social, academic, and emotional growth. Hence these disorders need to be identified and rectified at the earliest. Along with this speech therapists are not available in remote areas. Currently the evaluation methods used—such as manual assessments used by speech-language pathologists (SLPs) and offline clinical assessments—are biased, time-consuming, and less accessible in remote areas. With the dawn of Artificial intelligence and deep learning techniques, it is only fair to integrate it in speech therapy. Hence, AutoDEAP, a deep learning based, specifically a CNN based model which is optimized with transformer-based temporal modeling and metadata fusion is proposed. The model is trained on annotated speech samples using audio features like Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms to take manual speech misarticulation assessment to an advanced level. This makes early and remote detection possible, improving access to therapy, supporting SLPs in quick decision-making, and accelerating timely interventions. This has the potential to transform pediatric speech therapy by improving accessibility, diagnostic precision, and therapeutic results, ultimately helping children with speech misarticulation in their overall development and their involvement in educational and social environments.

Keywords: CNN, misarticulation assessment, speech language pathology, pediatric speech therapy, MFCCs, Mel-spectrogram, transformer.

1. Introduction

Speech is how human beings express themselves, put-forth ideas thoughts and take stand in the society with confidence. Misarticulations in speech can take a blow on a person's confidence. Hence, there is a need to identify and rectify these articulations at an early age. To address this a digital application-AutoDEAP, to assess phonological and articulatory errors in children aged 5–10 years is developed. There are traditional methods, such as the Diagnostic Evaluation of Articulation and Phonology (DEAP) test where an SLP shows images to a child having misarticulation disorders and records the audio of the child- while he/she names the objects. These audio files are then

manually assessed by the Speech therapists. This process of manual assessment introduces subjective errors into the process (Broen et al., 2012). Hence, to alleviate these subjective, human errors there is a need for a system that would conduct the same tests in an automated format, ensuring standardization and decreasing observer bias. The application will be in accordance with widely recognized phonological and articulatory standards, taken for the Indian regions by including phonological norms, regional accents, and linguistic differences (Santosh Kumar & Sharma, 2024). The main target population includes children from linguistically diverse states.

This technology uses artificial intelligence and sophisticated speech technology to identify and categorize speech with accuracy (Choi et al., 2020; Benzeghiba et al., 2007). This study addresses the shortcomings of existing speech error assessment methods, which are often expensive, time-consuming, and not easily accessible to remote areas (Deka et al., 2024). It is designed to help and support speech therapists, educators, and parents by providing objective, datadriven feedback into a child's speech patterns, enabling timely and targeted speech-care.

2. Literature Review

This section provides an overview of recent research studies on the detection of speech disorders employing machine learning-based approaches. Different approaches have been proposed, and Convolutional Neural Networks (CNNs) were proposed to represent speech features and facilitate automatic disorder detection by authors Jothi K. R. et al., (2020); Kanimozhiselvi C. S. et al., (2021).The paper by Jothi K. R. et al., (2020) was also on automating speech assessment systems for evaluating speech impairments and although is relied on machine learning classifiers like SVMs and ensembles, it highlighted that deep learning techniques, especially CNNs could be used to achieve a better performance. The papers major gap lied in the lack of use of deep learning techniques and of standardized datasets. The paper by Kanimozhiselvi C. S. et al., (2021) was based on a mobile application develop to identify communication disorders in the Tamil language. This application was completely CNN based; however, they suggested of better CNN optimizations and a lack of a more diverse population validation.

ResNet was also proposed as a model to use for speech disorder detection in some papers. Kohlschein C. et al., (2017), employed pre-trained ResNet-34 for automatic classification of speech aphasia, which gave extremely good results on healthy speech dataset. But struggled when used with aphasic speech dataset, the study also mentioned data scarcity and generalizations as problems. Another paper that mentioned ResNet was that of Hamza. et al., (2023) which was a study on automatic detection of severity level of dysarthria. The author proposed Residual Network (ResNet) architecture to process short duration speech samples to detect the level of dysarthria, proving ResNet as a suitable model for speech data processing as the research resulted in high accuracy and F1 score. The paper concluded that ResNet has strong practical applicability. The gaps however were potential overfitting,short segment focus and the amount of resource requirement.

The significance of feature extraction has also been highlighted in the research study. Sidhu et al., (2024) shows the importance of Mel-Frequency Cepstral Coefficients (MFCCs) in the processing of speech and audio signals as compared to other feature extractors such as Perceptual Linear Predictions, wavelet-based features, and spectral features. It assesses prior work to highlight the robustness of MFCCs in the context of voice disorder detection. The paper by Liu et al. (2021) built a speech disorder classification system within the Automated Phonetic TranscriptionGading Tool (APTgt). This paper also highlighted the use of MFCC for feature extraction. The results validated the suitability of MFCC+DTW features for phonetics. However, there were certain gaps regarding exploration of deeplearning based embeddings and feature extractors. Further, Mohan Sharma et al. (2023) conducted a comparative analysis of feature extractors for speech disorders. They compared between five feature types: MFCCs, LPCCs (Linear Predictive Cepstral Coefficients), GFCCs (Gammatone Frequency Cepstral Coefficients), mel-filterbank energy and spectrogram representations. Their results proved that MFCC outperformed the others with high precision, recall and F1-scores while using small number of coefficients. The only limitation of the paper was detecting multiple non-fluent types occurring simultaneously within same speech samples.

Review studies by Joshy A. A. et al. (2022) and Hamza A. (2023) also suggested other models, including Support

Vector Machines (SVMs), Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM) networks. However, in comparison to CNN and ResNet architecture-based models, these alternate methods showed relatively poorer accuracy rates for the detection of disorders.

The literature review also included several other papers which lead to the observations mentioned in the following sections 2.1 and 2.2.

2.1 Key findings from the survey

- Feature Extraction Techniques: Most studies used conventional audio feature extraction methods, including Mel-Frequency Cepstral Coefficients (MFCCs), Gammatone Frequency Cepstral Coefficients (GFCCs), Spectral Centroid, Zero Crossing Rate (ZCR), and Mel-spectrograms (Abdul & Al-Talabani, 2022; Sidhu et al., 2024; Chu et al., 2009). These methods focus on the acoustic properties important for studying speech patterns, although their performance varies depending on the dataset and the type of speech disorder examined.
- Machine Learning and Deep Learning Models: Researchers have repeatedly used Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and ResNet-based architectures for classification tasks in speech disorder detection (Bhatt et al., 2021; Khan et

al., 2020; Choi et al., 2020). In situations where datasets were not sufficient, transfer learning approaches were often used to improve model performance (Chronopoulos et al., 2021). The reviewed studies generally show effective classification capabilities, even with limited training data.

- Dataset Challenges: A repetitive concern in all the papers is the lack of large, disorder-specific datasets, which restricts the model from generalizing across broader populations (Agarwal et al., 2020; Sharda et al., 2023). Poor audio quality and inconsistencies in data annotation further reduce system accuracy.
- Application-Specific Insights: Some research has focused on highly specific populations, such as individuals with Chinese aphasia or children with autism spectrum disorders (Gierut, 1986; Berti et al., 2020). While such targeted studies allow for disorder-specific assessment plans, their limited scope makes broader application tough.

2.2 Identified Gaps and Limitations

Many critical gaps were identified from the literature review:

- Dataset Limitations: A lot of the works depended on datasets that were either too small or too domainspecific, which can lead to reduced model performance and accuracy when applied to a broader audience (Agarwal et al., 2020; Sharda et al., 2023). For instance, models trained only on speakers of a specific language or on a similar group of students often fail to generalize well.
- Performance Constraints: Although CNN and ResNet models typically perform well in controlled environments, their working in real-world scenarios is sometimes disturbed by poor sound quality and the shortage of dialectal diversity in training corpora (Bhatt et al., 2021; Khan et al., 2020).
- Model Complexity: Some disorders involve speech sound differences so minute that even advanced models find it difficult to differentiate them properly (Broen, 2012).
- Underrepresented Disorders: Some speech disorders—such as aphasia in Mandarin speakers—have been studied less in a less detailed manner, reducing the need for broader coverage in both languages and conditions (Chronopoulos et al., 2021).

In conclusion, the literature survey for speech misarticulation detection shows CNN and ResNet both to be comparatively better models in recognizing speech patterns and therefore speech errors. It also shows how MFCC outperforms other feature extraction techniques and enhances the model peformance.

3. System Architecture- AutoDEAP System

This section discusses the architecture of the AutoDEAP system. Initially the system was trained on a word-level audio dataset of children speaking in English. This dataset was first extracted and preprocessed from the ASER dataset and then a dataset is created containing the preprocessed audio files after converting the audio files to MFCCs along with the metadata with word being the label for prediction. The metadata consists of the age-group and the location of the speaker. These MFCCs contain spatial, spectral, and temporal information for each audio files which better work for image classification models. 85% of the data is used for training and the remaining 15% is used for validation. The test data is generated by a Speech Language Pathologist along with the metadata.

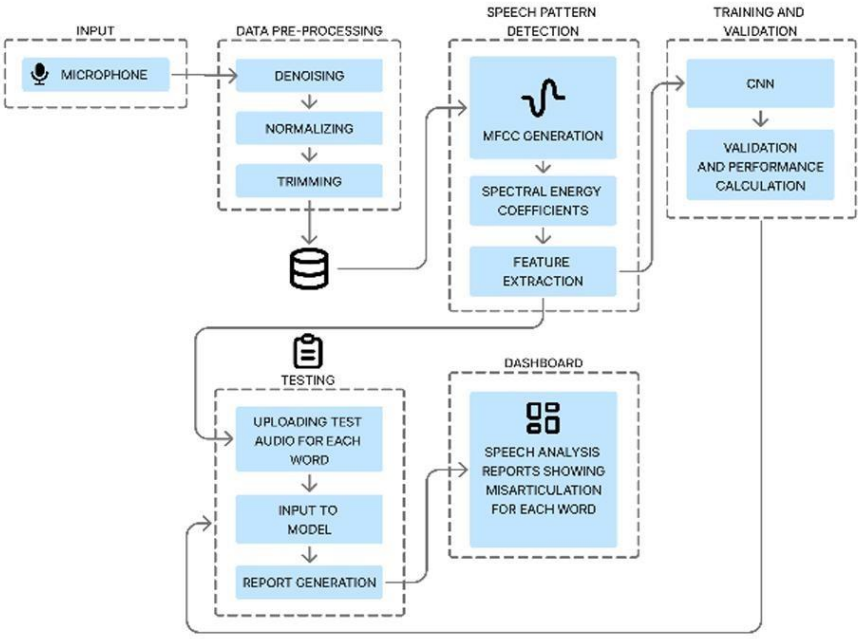


Figure 3.1 explains the architecture of our system which includes various components as follows:

- Input: Audio data is recorded with a microphone along with the metadata. The raw audio speech data is pre-processed prior to generating any features.
- Data Pre-Processing (for input and training): There are three general steps in outlining the pre-processing of the audio as follows:

• Denoising: Audio samples below an SNR (20db) or RMS (-53db) threshold are discarded, removing any background noise.
• Normalizing: Audio samples are normalized using a technique called peak normalization, wherein the waveform is divided by its maximum absolute amplitude so that the loudest point in the signal reaches a fixed target i.e. ±1. This ensures the audio uses the complete dynamic range without clipping, while the relative amplitude relationships between sample remains unchanged.
• Trimming: The audio samples undergo duration-based filtering of ~1 second in order to account for long audio samples. To remove unwanted silence/sound at the ends, clipping was performed with loudness thresholding.
• Feature Extraction: MFCC features with 20 coefficients are extracted from cleaned audio samples (both input and English word-level samples from the ASER dataset).
• Cleaned Dataset Generation: A new dataset is generated containing the metadata and the MFCC files (consisting a total of 7381 samples in our case).
• Training and Validation: The features extracted are used for training a custom CNN-Transformer hybrid with 85% of the cleaned dataset. The output of the transformer is concatenated along with the metadata embeddings as an input to another linear dense network, which directly results to the word labels. The outputs are validated with rest of the 15% of the dataset for 150 epochs.
• Testing: After training, the CNN-Transformer model makes a prediction based on the child's recorded audio, which is classified either as 'Misarticulated' or 'Normally Articulated'. The model outputs posterior probabilities over the vocabulary. For each iteration the model's best prediction is compared with the target word. If the word that is predicted aligned with the target word and the associated probability exceeds a certain confidence threshold, the articulation is classified as normally articulated. However, if the words are matched but the probability is below the predefined threshold the articulation was classified as misarticulated. In other case where the words are not aligned or matched at all the articulation was classified as misarticulated by default.

This architecture enables an end-to-end workflow from raw speech input to detailed misarticulation analysis, facilitating effective speech pattern recognition and assessment. To ensure reliability of the model, validation accuracy, validation loss, training loss and F1 score is observed. The training and validation loss-curves provide the information to detect overfitting or underfitting, thereby providing insight into the generalization capabilities of the model.

4. Proposed System

As was observed in the literature survey, ResNet and CNN proved to be the best deep learning techniques for speech data processing and speech disorder detection. In accordance with these findings the system went through a total of three iterations as follows.

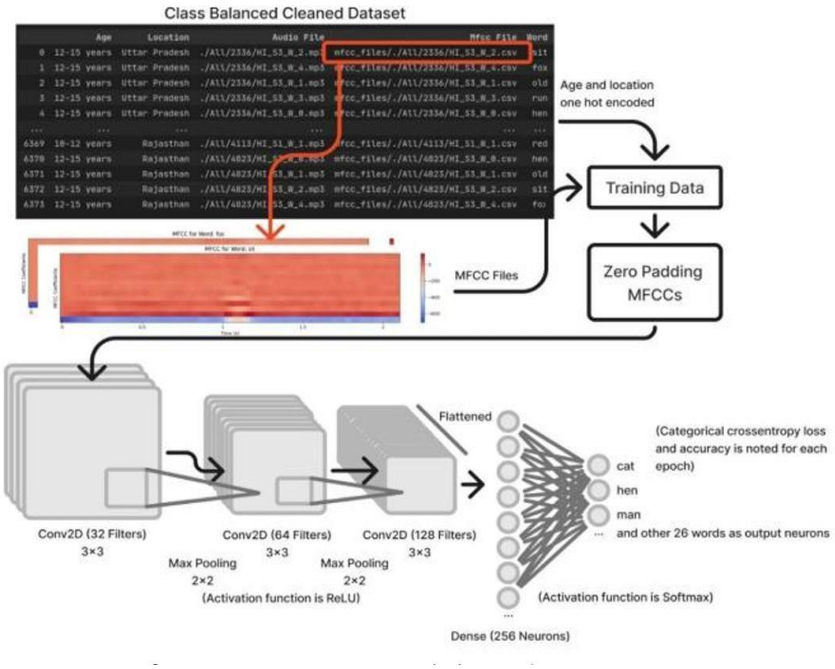
4.1 ResNet based workflow

The ResNet-based classifier for predicting articulation in children's speech accepts mel-spectrograms as input, which represent the time–frequency distribution of the audio signal. The spectrograms are processed by resizing to 224×224 pixels, converting to 3-channel RGB, and adding zero padding to have uniform input sizes. The network starts with a 7×7 convolution layer (64 filters), batch normalization, ReLU activation, and max pooling, followed by several residual blocks where stacked 3×3 convolutions are linked via skip connections to facilitate gradient flow and avoid vanishing gradients in deep networks. The residual learning is modeled by the equation,

$$y=F(x)+x, \quad \text{where } F(x)=W2\,\sigma(W1x+b1)+b2, \tag{4.8}$$

with σ(·) as the ReLU activation, making it stable for optimization even for very deep architectures. More residual blocks get hierarchical features (e.g., 128, 256 filters), and a global averaging pooling layer compresses spatial features into dense vectors. More metadata like age and location, one-hot encoded, may be concatenated to the pooled features for providing context information. Lastly, fully connected layers classify the outputs, with training conducted using optimizers such as Adam or SGD and learning rate schedulers to enhance convergence, with predictions being verified using confidence thresholds to discriminate between "Normal" and misarticulated speech.

4.2 CNN based workflow

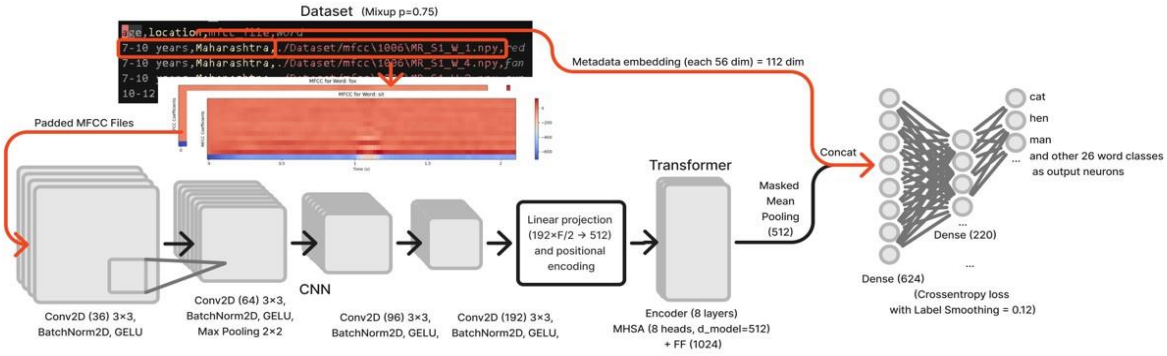


In figure 4.2.1, The CNN model for child speech articulation pattern classification has MFCCs derived from raw audio as input feature, retaining spectral and temporal information in the speech. The preprocessing includes high-frequency component augmentation by pre-emphasis, framing and Hamming windowing to constrain spectral leakage, and zeropadding for normalization of MFCC dimensions. Metadata such as age and location are one-hot encoded and added for context awareness. The CNN model first has the convolutional layers (32, 64, 128 filters of 3×3) extracting hierarchical and local features from the MFCCs sequentially, with max pooling (2×2) layers decresing spatial dimensions but maintaining the key features. This is mathematically represented by the convolution operation:

$$(W(l) * X(l-1))i,j = \sum Ku=-01 \sum Kv=-01 Wu(,lv)Xi(+l-u1,j)+v + b(l), \quad (4.3)$$

where the filters learn to recognize patterns such as phoneme transitions or articulation signals. The feature maps extracted are flattened and passed through dense layers (e.g., 256 neurons) before being input to the output layer, where a Softmax activation calculates class probabilities for 26 categories of speech. The network is trained using categorical cross-entropy loss with L2 regularization to minimize prediction errors and prevent overfitting to be able to correctly classify articulation errors in children's speech.

4.3 Optimized CNN- AutoDEAP Model



In Figure 4.3.1, From Figure 4.3.1, we observe the CNN–Transformer hybrid network employed for classifying articulation patterns of children's speech. The network is a combination of CNN-based local

feature extraction from MFCCs and Transformer-based global sequence modeling with the incorporation of metadata features (age and location). The dataset is built from CSV files associating each sample with its MFCC file, spoken word label, and speaker metadata (age, location). Classes are balanced at training time with a Weighted Random Sampler to provide an unbiased learning environment. Metadata is embedded where location and age are projected into integer indices and embedded into 56-dimensional vectors each (Config: META_EMB = 56). These embeddings are then concatenated with learned audio features prior to classification. Each sample is a 2D array of MFCCs (frames × coefficients), which are transposed, zero-padded to match length, and batched into batches using PyTorch's pad_sequence.

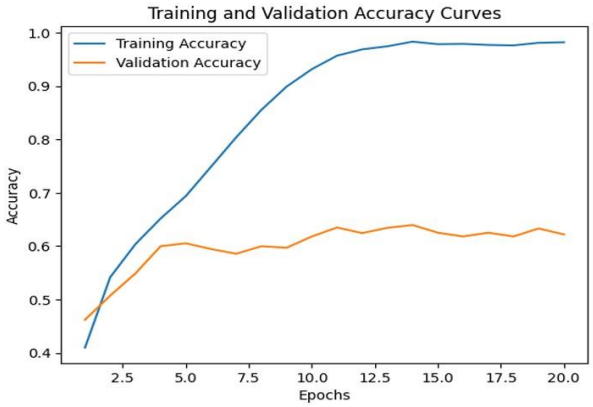
The CNN front-end transforms MFCC sequences into higher-level maps of acoustic features by way of successive convolutional layers: Conv Layer 1 (in=1, out=36, kernel=3, padding=1) → BatchNorm(36) → GELU;

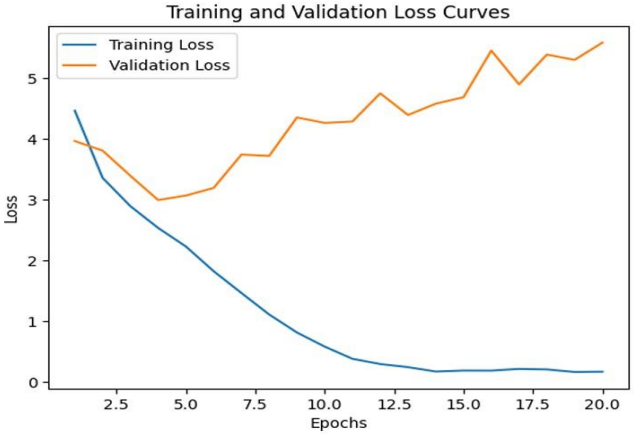
Conv Layer 2 (36→64, kernel=3, padding=1) → BatchNorm(64) → GELU; MaxPooling with 2×2 pooling; Conv Layer 3 (64→96, kernel=3, padding=1) → BatchNorm(96) → GELU; and Conv Layer 4 (96→192, kernel=3, padding=1) → BatchNorm(192) → GELU. The CNN output embedding dimension is 192 channels × downsampled frequency bins, which are projected to 512-dimensional vectors (Config: TRFM_D_MODEL = 512). A learnable positional encoding tensor of shape [1, 1000, 512] is added to the embeddings of CNN. The Transformer has 8 encoder layer stacks (Config: TRFM_LAYERS = 8) each with model size 512, multi-head self-attentive with 8 heads (TRFM_NHEAD = 8), feed-forward size 1024 (TRFM_FF_DIM = 1024), and dropout of 0.35 (DROPOUT = 0.35). LayerNorm is used before attention/FFN, with residuals plus dropPath stochastic depth (drop_prob=0.1). Masked mean pooling is applied over valid time steps to get a fixed-size sequence embedding.

For the integration of metadata, the age embedding has 56 dimensions and the location embedding also has 56 dimensions, so the final fused embedding dimension size = 512 (Transformer) + 56 (Age) + 56 (Location) = 624. The classification head is LayerNorm(624), Dropout(0.35), Linear(624 → 220), GELU, Dropout(0.35), and Linear(220 → n_classes) with Softmax at inference. The training setup involves AdamW optimizer (lr = 2.5e-4, weight decay = 1e-2), WarmupCosineScheduler with 15 epochs of warmup then cosine annealing down to min lr = 1e-6, and training for 150 epochs with batch size 32. The loss is CrossEntropy with label smoothing = 0.12, with mixup augmentation at probability 0.75 and α = 0.4. Gradient clipping is initialized with max-norm = 3.0, and DropPath regularization is used in Transformer layers with 15% of the dataset being used for validation.

5. Results and Discussions

5.1 Results – ResNet Model





Training accuracy increases steadily towards 1.0, indicating excellent fit on the training set. Validation accuracy ultimately levels of around 0.6 (and oscillates), further indicating lack of generalization. Loss curves are telling: training loss decreases steadily while validation loss begins to aim after an initial decrease, supporting the assertion of overfitting (He et al., 2016).

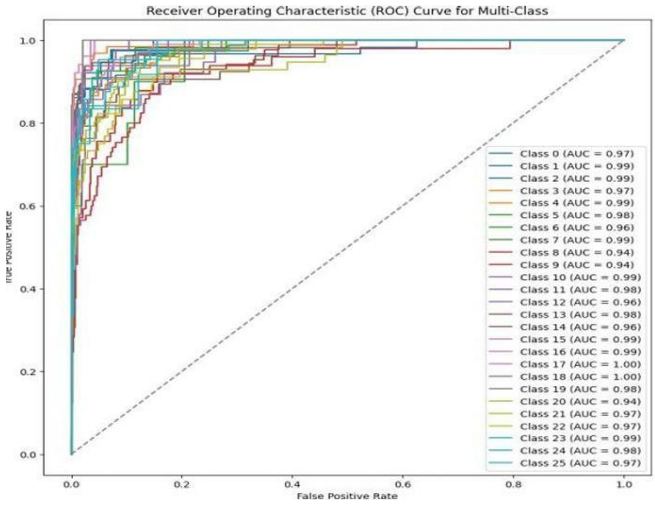
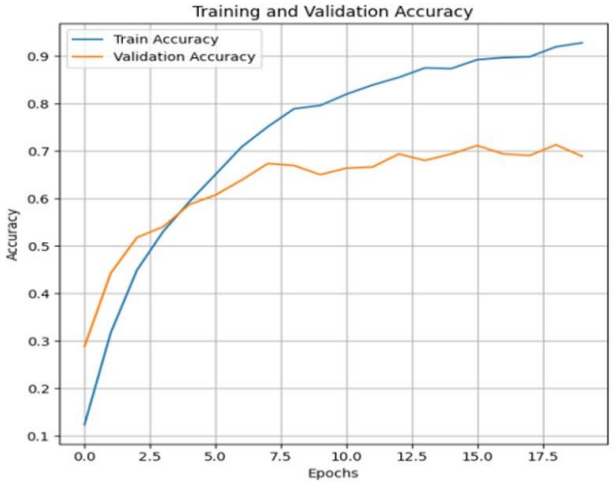
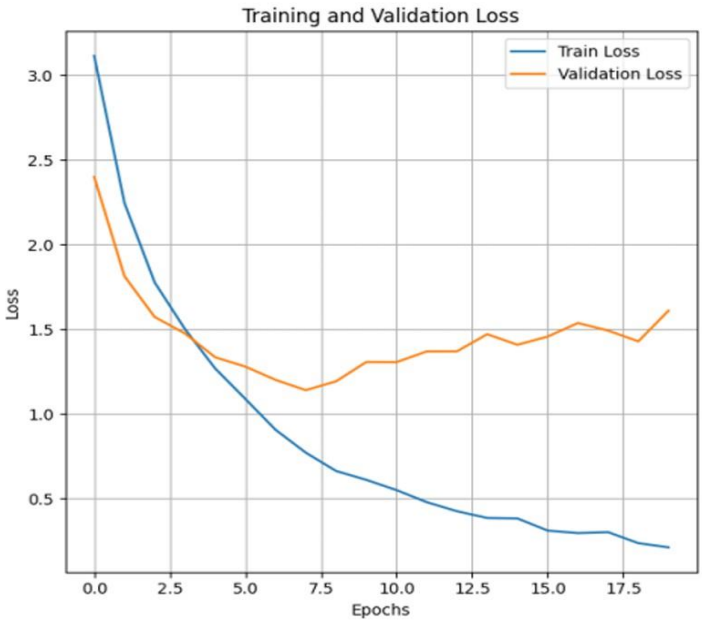


Figure 5.1.3 displays the ROC curves for the ResNet model. Most classes are at or near AUC values of 0.94 to 1.0, with classes 17 and 18 having perfect separation. The clustering of the curves shows that most classes are in the upper left part of the graph, indicating high TPR and low FPR (Sharda et al., 2023) With comparison to fitted models, the CNN had the better validation accuracy of 71.33% against ResNet's 63.48%. Even if ResNet is a deeper model with residual connections to combat vanishing gradients, it did not surpass the CNN accuracy. However, ResNet did have more stable loss curves (indicating more stable or smoother learning). Overall, while both models displayed overfitting, for this problem CNN was better than ResNet, as far as the generalization error and accuracy of the models. Adding data augmentation, dropout, and hyperparameter tuning might improve generalization (Bhatt et al., 2021; Khan et al., 2020). Fixing the learning rate schedule and increasing the number of epochs can also result in further improvement.

Therefore, in comparison, even though it had a deeper structure and residual connections incorporated too, ResNet only attained a validation accuracy of 63.48% after 20 epochs. While it was less than what was achieved by the CNN model, it still suggests a positive learning ability from complex patterns in the MFCCs. The skip connections helped to combat the vanishing gradients issue that plagued other deep learning models which resulted in a more stable learning procedure for the reasonable model. The training and validation loss curves of ResNet were more stable towards a decline in comparison to CNN, i.e. it is noted that there is not as much variation therefore is a more stable learning process. While the model may be rated as the lesser accurate model, based on the findings, the system made available the ability for more stable and consistent learning. Overall, the CNN model outperforms the ResNet in accuracy as well as generalization. The area of improvement includes the overfitting that is occurring with the CNN model. The common areas that could be improved in it include data augmentation, dropout methods and hyperparameter tuning like the learning rate and number of epochs. In overcoming the overfitting issue and modifying these parameters to optimally tuned values, I should be able to even produce a more compelling model overall.

5.2 ***Results – CNN Model***





Figures 5.2.1 and 5.2.2 compare the performance of the CNN model across 20 epochs, including training and validation accuracy and loss. From the accuracy chart, we see that one can observe both training accuracy consistently going up, exceeding 90%, and validation accuracy stabilizing at about 70%. There is a clear separation between training accuracy and validation accuracy, and further evidence of overfitting, since the model seems to be learning more and more with respect to training-only data, but again not performing well in terms of generalization, poor performance on unseen data. The pattern of the loss supports this. Indeed, training loss consistently goes down, suggesting increasing learning, but validation loss increases after some small improvements. This suggests that the generalization of the model on unseen data decreases over time (Bhatt et al., 2021; Khan et al., 2020).

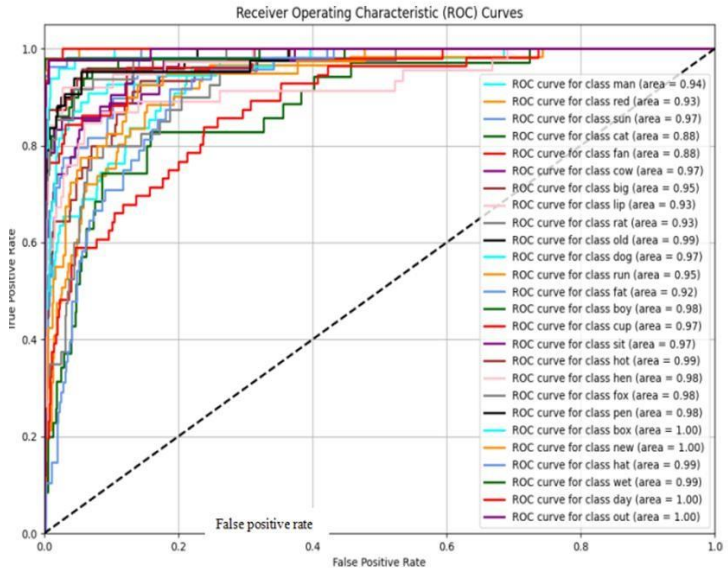
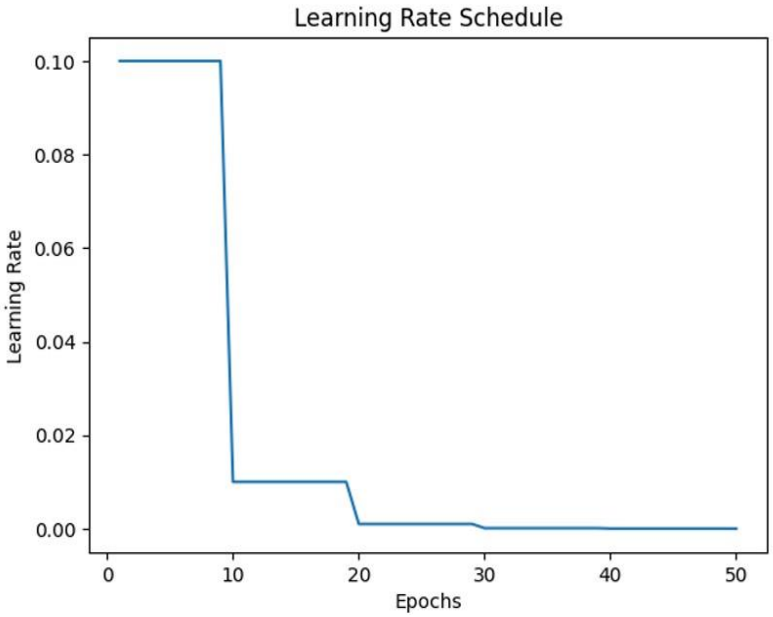
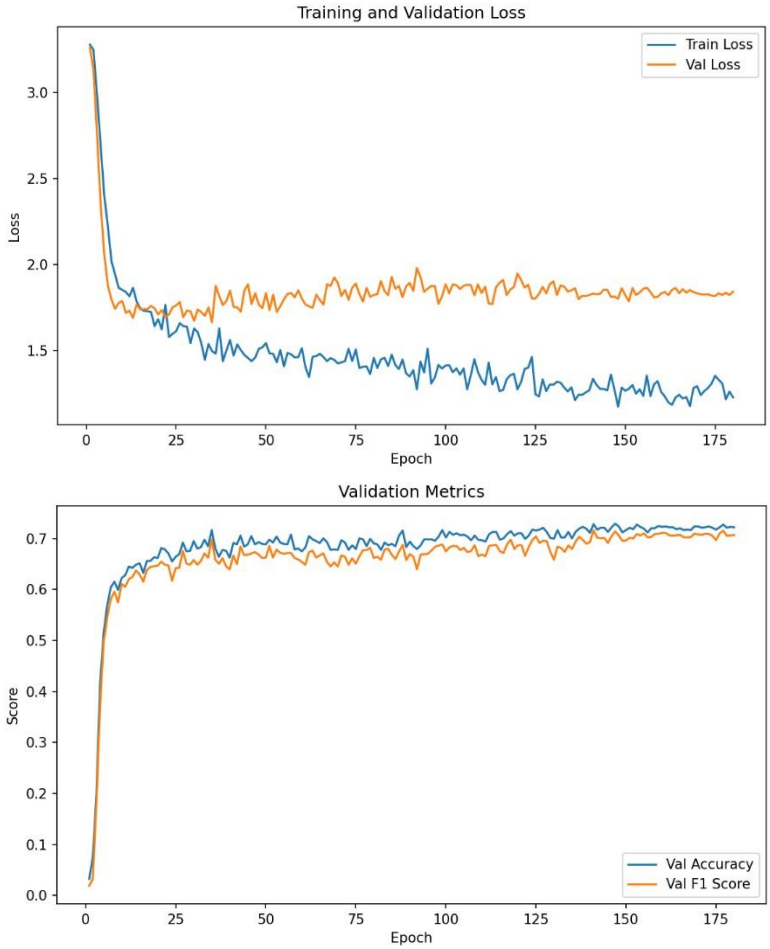


Figure 5.2.3 shows ROC curves for every class within the multi-class classification problem. Each curve demonstrates the trade-offs between false positive rate (FPR) and true positive rate (TPR), with AUC (Area Under the Curve) noted for each class. High values of AUC close to 1.0 show high discriminative power, and many classes have a perfect AUC of 1.0. Values of AUC slightly below 1.0 highlight possibilities for model improvement. The diagonal reference line shows random classification performance (Sharda et al., 2023).



The learning rate schedule in Figure 5.2.4 indicates step-wise decay, starting with an initial rate of 0.1 for greater updates during initial training, followed by sharp drops at epochs 10, 20, and 30. This slow reduction of the learning rate enables the optimizer to settle training and prevent overshooting the optimal solution (Kingma & Ba, 2014).

5.3 ***Results – Optimized CNN Model***



Figures 5.3.1 and 5.3.2 compare the performance of the CNN-Transformer model across 150 epochs, including training and validation loss and validation accuracy and f1 score. From the accuracy chart, we see that one can observe both validation accuracy and f1 score consistently going up, crossing 70%, and validation accuracy stabilizing at about 72.90%. There is a clear separation between training accuracy and validation accuracy, since the model seems to be learning more, it also performs well in terms of generalization, on unseen data. The training loss consistently goes down, suggesting increasing learning.

The model gives us a f1 score of 71.36% and an accuracy of 72.90% which is an increase from 71.33% obtained from the CNN-based model.

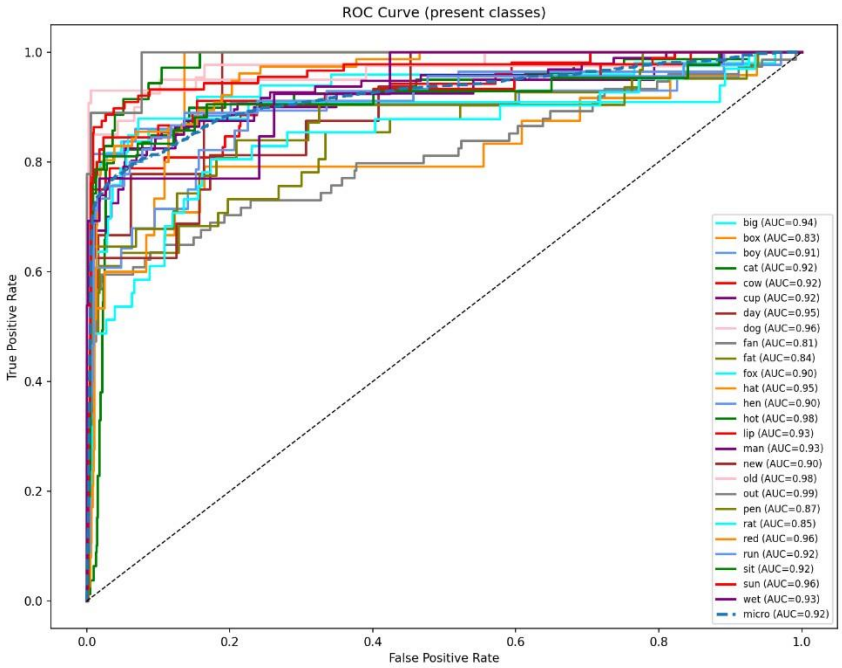


Figure 5.3.3 shows ROC curves for every class within the multi-class classification problem. Each curve demonstrates the trade-offs between false positive rate (FPR) and true positive rate (TPR), with AUC (Area Under the Curve) noted for each class. High values of AUC close to 1.0 show high discriminative power, and many classes have a perfect AUC of 1.0. Values of AUC slightly below 1.0 highlight possibilities for model improvement. The diagonal reference line shows random classification performance (Sharda et al., 2023).

6. Conclusion

In conclusion, this research demonstrates the effectiveness of the optimized CNN-Transformer based hybrid model in detecting misarticulations in children's speech using Mel-spectrogram features and MFCCs. This optimized CNN model works better than the ResNet architecture and the plain CNN architecture in terms of accuracy and speech pattern recognition. While there was slight-overfitting observed, feature engineering methods such as data augmentation, dropout, weight decay, and optimized hyperparameters improve the performance. The research makes use of a hybrid architecture to achieve maximum accuracy. Real-time deployment of this model benefits speechlanguage pathologists with immediate, child-specific feedback, making therapy services accessible and efficient.

7. Future Scope

The progression of this project relies on its effectiveness in improving assessment and intervention for speech misarticulation. One possible direction is to implement more sophisticated speech recognition technologies that can adapt to a variety of accents and dialects so that it is accessible to speakers of other linguistic profiles (Xiong et al., 2018). If prosodic features such as rhythm, pitch, and stress patterns could also be included, this would increase the effectiveness of the model in detecting articulation errors (Kane & Gobl, 2017).

Interactive features that provide real-time feedback and gamify therapy tasks could also increase user engagement, especially with children; however, this aspect can also be appropriate for adults with speech related disabilities (Chen et al., 2007). Adding a teletherapy component could also broaden access to speech therapy (Cason & Brannon, 2011).

To improve monitoring and tracking progress using continuous speech, new and existing wearable technologies and IoT-enabled devices for the purposes of monitoring and tracking (Patel et al., 2012). Partnering with speech-language pathologists and educators would result in more individualized therapy programs. Additionally, gaining access to many more regional languages and speech areas would improve the diversity of the dataset and resource for the model (Gauthier et al., 2016). Together, all of these changes can help this system become an appropriate tool to improve communication skills on a global level.

Acknowledgement

We would like to express our extreme gratitude to the Department of Artificial Intelligence and Machine learning, Dwarkadas Jivanlal Sanghvi College of Engineering for providing us with the knowledge and opportunity to carry this research.

We would also like to thank SLP Ramya Adiseshan for her immense support in helping us understand the field of speech therapy and thus in applying our knowledge to the field.

Author Contribution

All the authors contributed equally to the conceptualization, literature survey, methodology analysis and writing of this research. Each author has read and approved the final manuscript.

Statements and Declaration

Ethical Statement

This study does not contain any studies with animal or human subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Funding Statement

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data Availability

The data used to train the CNN and Resnet Models was the ASER Dataset taken from Kaggle.If there is a requirement to provide data we will provide it with immediate effect.

References

1. Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access, 10*, 122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444
2. Agarwal, D., Gupchup, J., & Baghel, N. (2020). A dataset for measuring reading levels in India at scale. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9210–9214.
3. https://doi.org/10.1109/ICASSP40776.2020.9053380
4. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006
5. Berti, L., Guilherme, J., Esperandino, C., & de Oliveira, A. (2020). Relationship between speech production and perception in children with speech sound disorders. *Journal of Portuguese Linguistics, 19*(1), 13.
6. https://doi.org/10.5334/jpl.244
7. Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics, 10*(20), 2470.
8. https://doi.org/10.3390/electronics10202470
9. Broen, P. A. (2012). Patterns of misarticulation and articulation change. In *Speech and Language* (Vol. 8). Elsevier. https://doi.org/10.1016/B978-0-12-608608-9.50008-5
10. Chen, Y.-J., Huang, J.-W., Yang, H. M., Lin, Y.-H., & Wu, J.-L. (2007). Development of articulation assessment and training system with speech recognition and articulation training strategies selection. In *ICASSP-88: 1988 International Conference on Acoustics, Speech, and Signal Processing* (Vol. 4, pp. IV-209–IV-212). https://doi.org/10.1109/ICASSP.2007.367200
11. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology, 9*(2), 14.
12. https://doi.org/10.1167/tvst.9.2.14
13. Chronopoulos, S. K., Schmitt, M., Schüller, B., Jeschke, S., & Werner, C. J. (2021). Exploring speech language therapy through information communication technologies, machine learning and neural networks. *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 193–198.
14. https://doi.org/10.1109/ISMSIT52890.2021.9604553
15. Chu, S., Narayanan, S., & Kuo, C.-C. J. (2009). Environmental sound recognition with time–frequency audio features.
16. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1142–1158.
17. https://doi.org/10.1109/TASL.2009.2017438
18. Deka, C., Shrivastava, A., Abraham, A. K., Nautiyal, S., & Chauhan, P. (2022). AI-based automated speech therapy tools for persons with speech sound disorders: A systematic literature review [Preprint].
19. https://doi.org/10.21203/rs.3.rs-1517404/v2
20. Deka, C., Shrivastava, A., Abraham, A. K., Nautiyal, S., & Chauhan, P. (2024). AI-based automated speech therapy tools for persons with speech sound disorder: A systematic literature review. *Speech, Language and Hearing*, 1–22. https://doi.org/10.1080/2050571X.2024.2359274
21. Durak, L., & Arikan, O. (2003). Short-time Fourier transform: Two fundamental properties and an optimal
22. implementation. *IEEE Transactions on Signal Processing, 51*(5), 1231–1242.
23. https://doi.org/10.1109/TSP.2003.810293
24. Estévez, D., Terrón-López, M.-J., Velasco-Quintana, P. J., Rodríguez-Jiménez, R.-M., & Álvarez-Manzano, V. (2021). A case study of a robot-assisted speech therapy for children with language disorders. *Sustainability, 13*, 2771. https://doi.org/10.3390/su13052771
25. Gierut, J. A. (1986). Sound change: A phonemic split in a misarticulating child. *Applied Psycholinguistics, 7*(1), 57–
26. 68. https://doi.org/10.1017/S0142716400007189
27. Hamza, A., Addou, D., & Kheddar, H. (2023). Machine learning approaches for automated detection and classification of dysarthria severity. 2023 *2nd International Conference on Electronics, Energy and Measurement (IC2EM),* 1–6. https://doi.org/10.1109/IC2EM59347.2023.10419588
28. Joshy, A. A., & Rajan, R. (2022). Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, 1147–1157. https://doi.org/10.1109/TNSRE.2022.3169814
29. Jothi K. R., & Mamatha V. L. (2020). A systematic review of machine learning based automatic

speech assessment system to evaluate speech impairment. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*,
30. 175–185. https://doi.org/10.1109/ICISS49785.2020.9315920
31. Karunasekara, P., Deshitha, S., De Alwis, D., Karunarathna, D., Lokuliyana, S., & Gamage, N. (2023). Children's speech disorders identification and therapy treatment. *2023 5th International Conference on Advancements in*
32. *Computing (ICAC)*, 179–184. https://doi.org/10.1109/ICAC60630.2023.10417158
33. Kanimozhiselvi C. S., & Santhiya, S. (2021). Communication disorder identification from recorded speech using machine learning assisted mobile application. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 789–793. https://doi.org/10.1109/ICICV50876.2021.9388493
34. Kim, G., Eom, Y., Sung, S. S., Ha, S., Yoon, T.-J., & So, J. (2024). Automatic children speech sound disorder detection with age and speaker bias mitigation. *Proc. Interspeech 2024*, 1420–1424. https://doi.org/10.21437/Interspeech.20241799
35. Kim, H., Martin, K., Hasegawa-Johnson, M., & Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clinical Linguistics & Phonetics, 24*(10), 759–770. https://doi.org/10.3109/02699206.2010.497238
36. Khan, A., Sohail, A., Zahoora, U., & Others. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review, 53*, 5455–5516. https://doi.org/10.1007/s10462-020-09825-6
37. Kohlschein, C., Schmitt, M., Schüller, B., Jeschke, S., & Werner, C. J. (2017). A machine learning based system for the automatic evaluation of aphasia speech. *2017 IEEE 19th International Conference on e-Health Networking,*
38. *Applications and Services (Healthcom)*, 1–6. https://doi.org/10.1109/HealthCom.2017.8210766
39. Liakina, N., & Liakin, D. (2023). Speech technologies and pronunciation training: What is the potential for efficient corrective feedback? In U. Kickhöfel Alves & J. Alcantara de Albuquerque (Eds.), *Second Language Pronunciation:*
40. *Different Approaches to Teaching and Training* (pp. 287–312). De Gruyter Mouton.
41. https://doi.org/10.1515/9783110736120-011
42. Liu, J., et al. (2021). Speech disorders classification in phonetic exams with MFCC and DTW. *2021 IEEE 7th*
43. *International Conference on Collaboration and Internet Computing (CIC)*, 35–40.
44. https://doi.org/10.1109/CIC52973.2021.00015
45. Malakar, M., & Keskar, R. B. (2021). Progress of machine learning based automatic phoneme recognition and its prospect. *Speech Communication, 135*, 37–53. https://doi.org/10.1016/j.specom.2021.09.006
46. Melle, N., Lahoz-Bengoechea, J. M., Nieva, S., & Gallego, C. (2023). Temporal acoustic properties of the sibilant fricative /s/ for the differential diagnosis of dysarthria and apraxia of speech in Spanish speakers. *Clinical Linguistics & Phonetics, 38*(9), 838–856. https://doi.org/10.1080/02699206.2023.2244646
47. Menyuk, P. (1980). The role of context in misarticulations. In *Child Phonology*. Academic Press. https://doi.org/10.1016/B978-0-12-770601-6.50016-4
48. Mohan Sharma, V., Kumar, V., Mahapatra, P. K., & Gandhi, V. (2023). Comparative analysis of various feature extraction techniques for classification of speech disfluencies. *Speech Communication, 150*, 23–31. https://doi.org/10.1016/j.specom.2023.04.003
49. Namasivayam, A. K., Coleman, D., O'Dwyer, A., & van Lieshout, P. (2020). Speech sound disorders in children: An articulatory phonology perspective. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.02998
50. Naz, A., Noor, H., Hussain, A., Bukhari, S., Pervaiz, N., & Inam, I. (2023). Comparison of traditional articulation therapy and picture articulation test in children with articulation disorders. *Journal of Health and Rehabilitation*
51. *Research, 3*(2), 448–453. https://doi.org/10.61919/jhrr.v3i2.139
52. Prasetio, B. H., Yusuf, D. O., Syauqy, D., & Chilmi, S. (2024). Spectral gating for noise reduction in speech stress recognition system. *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 149–155. https://doi.org/10.1109/IAICT62357.2024.10617782
53. Santosh Kumar, S. S. B., & Sharma, M. (2024). Development of articulation test in Garhwali language: A preliminary study. *An International Journal of Otorhinolaryngology Clinics, 16*(1), 5–7. https://doi.org/10.5005/jp-journals10003-1500
54. Sharda et al. (2023). A systematic review of online speech therapy systems for intervention in childhood speech communication disorders. *MDPI, 12*(4), 567–588. https://doi.org/10.3390/s22249713
55. Sidhu, M. S., Latib, N. A. A., & Sidhu, K. K. (2024). MFCC in audio signal processing for voice disorder: A review.
56. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19253-1
57. Smith, L. K., et al. (2022). Preschooler processing of common versus rare misarticulations. *Journal of Speech and Language Processing, 18*(3), 204–212. https://doi.org/10.1044/2017_JSLHR-S-16-0379
58. Xu, Y. (2024). Image classification based on ResNet models. *Science and Technology of*

Engineering, Chemistry and Environmental Protection, 1. https://doi.org/10.61173/c5xnwg67

59. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90
60. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980
61. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929–1958.
62. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. *Proceedings of the International Joint Conference on Neural Networks*, 2047–2052.
63. https://doi.org/10.1109/IJCNN.2005.1556215
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
65. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering,*
66. *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191
67. Cason, J., & Brannon, J. A. (2011). Telehealth in speech-language pathology: The use of telepractice to deliver speechlanguage services to school-age children. *International Journal of Telerehabilitation, 3*(1), 19–28.
68. https://doi.org/10.5195/ijt.2011.6071
69. Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation, 9*(21), 1–17.
70. https://doi.org/10.1186/1743-0003-9-21