ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

Soft Palatal Trainer with Visual Feedback for Children with Cleft Palate

Yiyan Pei

Dunn School, Los Olivos, United States of America

Abstract: *Children with cleft palate often experience persistent speech difficulties caused by abnormal oral structures and impaired velopharyngeal function. Conventional pronunciation training relies on repeated practice and professional feedback, yet children practicing independently can hardly judge whether their articulatory movements are appropriate. To address this problem, this study proposes a soft-palatal training system that combines a hardware training device with an artificial intelligence-based visual feedback module. The hardware part is a silicone oral appliance driven by fishing lines and pneumatic airbags, which can bend bidirectionally to guide the soft palate into a correct position. The software part extracts facial landmarks from speech videos and computes geometric, temporal, and statistical features to characterize mouth movements. A hybrid one-dimensional convolutional neural network, bidirectional gated recurrent unit, and Transformer architecture is developed to classify four predefined visual pronunciation patterns. The GRID audiovisual speech corpus is used to develop and evaluate the proposed model. The model achieves an accuracy of 87.30%, a macro F1-score of 87.43%, a precision of 87.38%, and a recall of 87.49%. It outperforms the LSTM, GRU, Transformer, CNN, and MLP baselines. These findings support the feasibility of the proposed visual feedback approach for pronunciation analysis and its potential use in home-based articulation training for children with cleft palate.*

Keywords: *cleft palate; visual speech recognition; pronunciation training; visual feedback; deep learning*

1. Introduction

Cleft lip and/or palate is a common congenital craniofacial condition caused by incomplete fusion of facial structures during fetal development. In the United States, cleft lip with or without cleft palate occurs in approximately 10 per 10,000 live births [1]. Children with cleft palate may experience difficulties with speech, hearing, feeding, and dental development. Speech disorders are particularly common and may include hypernasality, nasal air emission, and articulation errors associated with velopharyngeal dysfunction [2]. The motivation for this work is personal: the author's younger brother and uncle both have cleft lip and palate. Their daily communication difficulties prompted the development of an affordable, accessible training device.

Current approaches to treatment and training include mobile speech therapy applications, professional articulation therapy, and surgery. Mobile applications such as Smile Train provide pronunciation exercises through games and stories but lack real-time feedback on the patient's mouth movements. Online courses, including those offered by Teachers College, Columbia University, train professionals to deliver articulation therapy. The therapy therefore depends heavily on specialist support, while the courses themselves do not help patients control their soft palate position independently. Surgery to improve speech is another option, but it can be expensive and uncomfortable and does not guarantee complete correction. A low-cost system offering real-time, quantitative, and personalized feedback is therefore still needed.

This study proposes a soft palate training system that combines a hardware training device with an AI-based visual feedback module. A line-driven silicone device with an airbag is designed to adjust the position of the soft palate. The visual analysis module analyzes the patient's mouth movements during pronunciation and produces scores that provide clear, concise feedback during training.

Visual speech recognition (VSR) extracts speech-related information from visible movements of the face and lips. Earlier VSR methods relied largely on handcrafted appearance and geometric features, whereas more recent approaches use deep neural networks to learn spatial and temporal representations

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

directly from speech videos [3]. Stafylakis and Tzimiropoulos combined spatiotemporal convolution, residual networks, and bidirectional LSTM networks for word-level lipreading [4]. Petridis et al. later used residual networks together with bidirectional gated recurrent units to model audiovisual speech sequences [5]. The Transformer architecture provides another approach to sequence modeling by using self-attention to capture relationships between different temporal positions [6]. Most of these studies aim to identify words, phonemes, or complete utterances. Pronunciation training has a somewhat different goal: instead of reconstructing the full speech content, it is more useful to describe how visible articulatory patterns change during pronunciation and to use this information to generate feedback. Based on this idea, this study develops a hybrid 1D-CNN-BiGRU-Transformer architecture to classify pronunciation segments into four project-defined visual movement categories: vertical, horizontal, radial, and baseline, using the GRID audiovisual speech corpus [7] for model development and evaluation.

2. System Overview

The proposed system consists of two parts that work together during rehabilitation training. The hardware part is a soft palatal trainer: a silicone oral appliance, actuated by fishing lines driven by a servo motor and supported by pneumatic airbags, gently bends to guide the patient's soft palate toward the correct position. A bending sensor measures the curvature of the device in real time so that the bending angle can be calibrated for each patient. The software part is a visual feedback module: the patient pronounces a target word in front of a camera, the system extracts mouth-shape features from the video, the trained deep learning model classifies the pronunciation pattern, and a score is returned to the patient so that the training process becomes visible and quantifiable. Because the device is simple and inexpensive, the system is suitable for both rehabilitation centers and home training. System logic flowchart is shown in Figure 1.

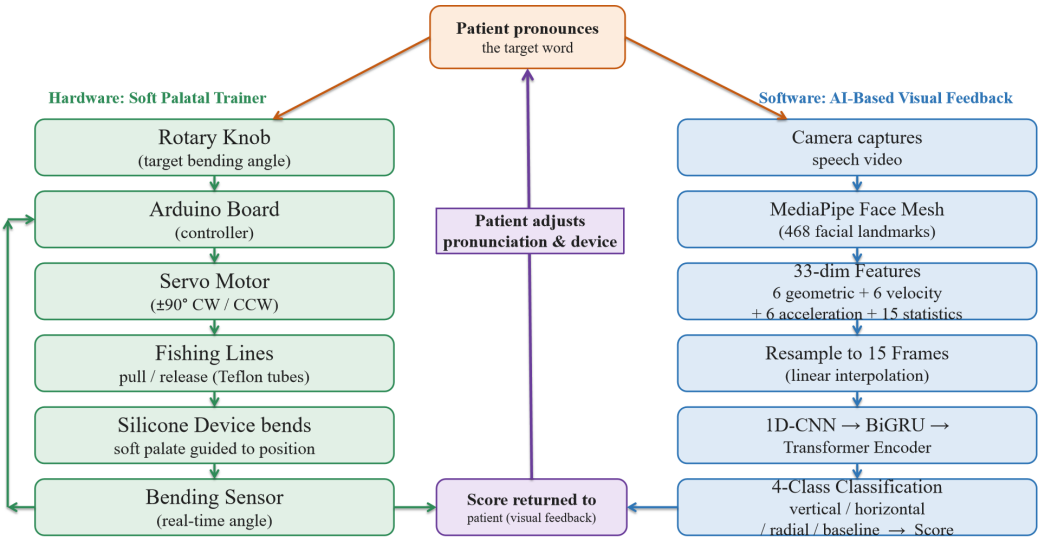


Figure 1: System Logic Flowchart

3. Hardware Design

The hardware of the soft palatal trainer includes a support structure, the silicone device itself, and an electronic control system. A 3D-printed bracket was designed to support the servo motor, and a second bracket with two holes holds the Teflon tubes that guide the fishing lines. A winding wheel with four holes connects the servo motor to the line-reel mechanism, and all components are mounted on a single base board, as shown in Figure 2.

Two generations of molds were designed for fabricating the silicone device. The first generation used a pneumatically actuated bending mechanism. The mold contains multiple small and thin air chambers; after the silicone is poured and cured, two silicone layers are bonded together with liquid silicone. When air is injected, the air chambers inflate and the device bends, as shown in Figure 3 and Figure 4. However, this generation can only bend in one direction and the bending angle is hard to control, which prevents personalized adjustment for each patient.

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

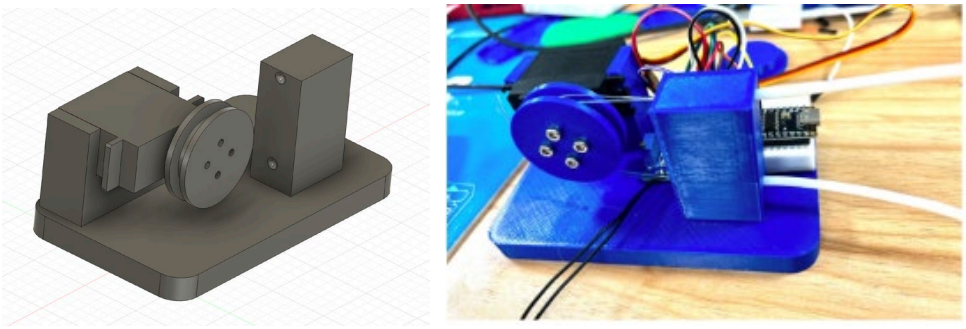


Figure 2: 3D-printed support bracket of the soft palatal trainer.

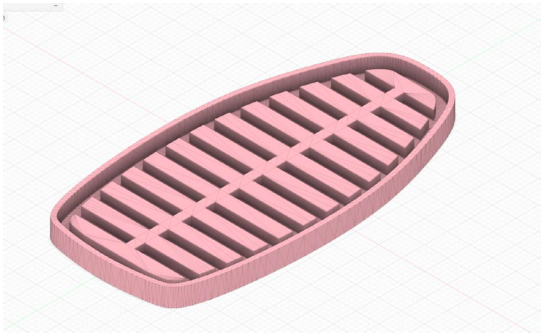


Figure 3: Mold design (version 1) for the pneumatic silicone device.

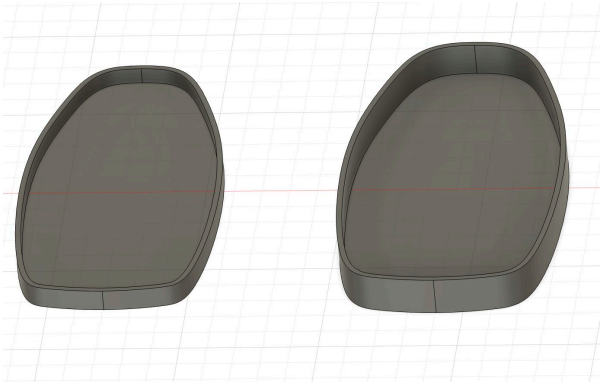


Figure 4: Air chambers and bonded cover plate of version 1.

The second generation adopts a line-driven mechanism inspired by wire-driven robotic arms, which enables compound bidirectional bending. The mold contains bilateral airbags that allow the device to anchor inside the oral cavity, and a thin airbag at the center of the mold occludes the cleft palate defect. Six holes on the mold wall allow fishing lines to pass through, and a thin support plate is placed inside the device to keep the hard palate region from bending. Only the soft palate section bends, as required for training. The mold of version 2 is shown in Figure 5.

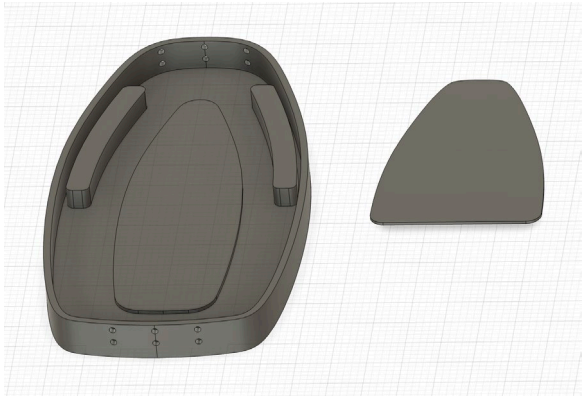


Figure 5: Mold design (version 2) for the line-driven silicone device.

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

An Arduino board controls the electronic system and connects to a bending sensor, a rotary knob, and a servo motor. The knob controls the servo, which can rotate 90° clockwise or counterclockwise to pull or release the fishing lines, bending the silicone device in the desired direction. The sensor measures the device's bending angle in real time. If the difference between the sensor reading and the knob setting exceeds a threshold, the servo turns in the appropriate direction to bend the device. Otherwise, it remains stationary. A winding wheel takes up and releases the lines, while Teflon tubes guide them along predetermined paths to keep the device's movement stable.

4. AI-Based Visual Feedback Method

The visual pronunciation analysis pipeline preprocesses the data, extracts facial landmarks, constructs features, and uses the CNN-BiGRU-Transformer model for classification.

4.1 Dataset and Preprocessing

This study used the GRID audiovisual speech corpus, which contains approximately 34,000 audiovisual sentence recordings from 34 speakers, each producing 1,000 sentences [7]. The original recordings were converted to MP4 format using FFmpeg. Available temporal alignment information was used to locate the start and end times of the target pronunciation segments. Samples with invalid timing information, including those whose end times preceded their start times, were removed during preprocessing.

MediaPipe Face Mesh was used to extract facial landmarks from each valid video frame. It represents facial geometry with a dense mesh of 468 landmarks [8]. Only landmarks needed to describe mouth movements and normalize facial measurements were retained: 13 and 14 for the inner lip; 0, 17, 61, and 291 for the outer lip; 1 for the nasal tip; and 33 and 263 for the eye corners. Frames without reliable landmark detections were excluded, as were pronunciation segments containing fewer than three valid frames, which provided insufficient temporal information for movement analysis.

Before feature extraction, mouth-related geometric distances were normalized by the interocular distance between landmarks 33 and 263 to reduce variation caused by facial size and camera distance. The processed pronunciation segments were grouped into four predefined visual movement categories: vertical, horizontal, radial, and baseline. These categories served as the class labels for model training and evaluation.

4.2 Feature Extraction

Six geometric features were calculated from the normalized facial landmarks: mouth aspect ratio, normalized vertical mouth distance, normalized horizontal mouth distance, inner-lip mouth aspect ratio, normalized nasolabial distance, and lip aspect ratio. These measurements describe mouth opening, lip width, inner-lip separation, and changes in the mouth's relative position during pronunciation.

Because static measurements cannot fully describe articulatory movements over time, first-order temporal differences were calculated for each of the six feature sequences to capture frame-to-frame changes in mouth configuration. Second-order differences captured changes in movement velocity, providing further information about articulation dynamics. To summarize the overall movement pattern, the mean, standard deviation, and maximum were calculated for the first five geometric features, yielding 15 sequence-level statistical descriptors.

The final representation contained 33 features: six geometric features, six first-order temporal differences, six second-order temporal differences, and 15 statistical descriptors. To accommodate differences in segment duration, the temporal sequences were resampled to a fixed length of 15 frames using linear interpolation before being passed to the model.

4.3 CNN-BiGRU-Transformer Model

The proposed model combines a 1D-CNN, a BiGRU, and a Transformer encoder to capture local movement patterns and temporal relationships in pronunciation sequences. The fixed-length feature representation first passes through two 1D convolutional layers, both with a kernel size of 3 and with 64 and 128 output channels, respectively. These layers extract short-term mouth movement patterns from neighboring frames.

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

The convolutional features then pass through a two-layer BiGRU with a hidden size of 128 in each direction. By processing the sequence both forward and backward, the BiGRU captures how articulatory movements develop across the full pronunciation segment. Its output feeds into a Transformer encoder with four attention heads and a model dimension of 256. Self-attention connects information across sequence positions and allows the model to give greater weight to frames that are more informative for classification.

Average pooling and max pooling summarize the Transformer output over time. The pooled features then pass through a two-layer classification head (256-64-4) with BatchNorm and Dropout. The head produces four output probabilities, one for each movement category: vertical, horizontal, radial, and baseline.

4.4 Model Training

The processed samples were randomly divided into training and validation sets at a ratio of 80:20. Stratified sampling was used to preserve the class distribution between the two subsets. Cross-entropy loss with a label-smoothing factor of 0.1 was used as the training objective. The model was optimized using AdamW with an initial learning rate of 0.001 and a weight decay of 0.05. ReduceLROnPlateau was used to adjust the learning rate when the validation loss stopped improving: if no improvement was observed for five consecutive epochs, the learning rate was reduced by half. Dropout values between 0.3 and 0.5 were used for regularization, and gradient clipping with a maximum norm of 1.0 was applied during training. Early stopping was set with a patience of 15 epochs, and the parameters corresponding to the lowest validation loss were retained. Five baseline models were used for comparison: LSTM, GRU, Transformer, CNN, and multilayer perceptron (MLP). Accuracy, macro F1-score, precision, and recall were used to evaluate classification performance. AI model training process is shown in Figure 6.

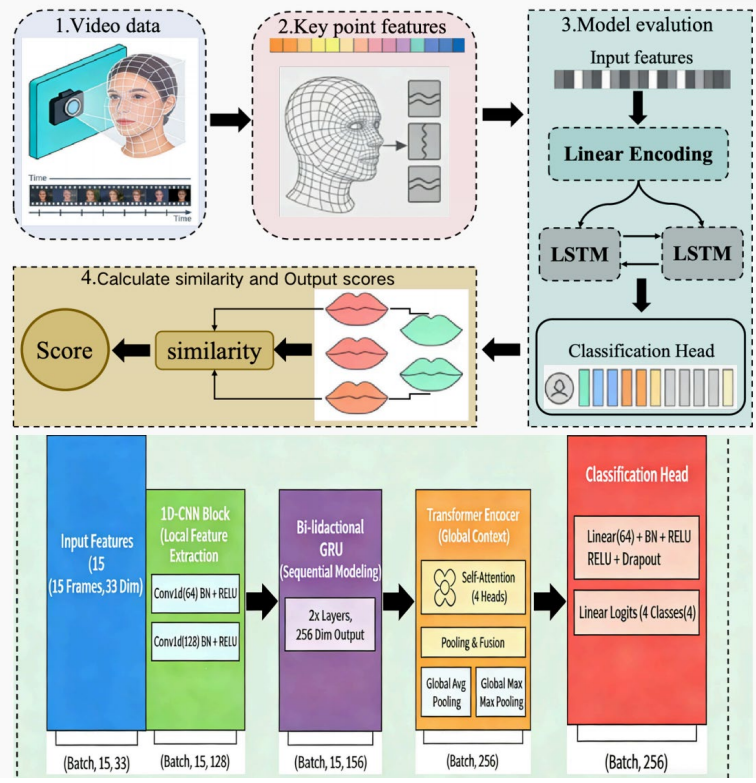


Figure 6: AI model training process

5. Experiments and Results

5.1 Visual Feedback Performance

The validation performance of the proposed CNN-BiGRU-Transformer model was compared with the five baseline models, and the results are summarized in Table 1. The proposed model obtained the

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

highest values for all four evaluation metrics, reaching an accuracy of 0.8730, a macro F1-score of 0.8743, a precision of 0.8738, and a recall of 0.8749. LSTM was the strongest baseline, with an accuracy of 0.8701 and a macro F1-score of 0.8713, while GRU achieved slightly lower results.

Table 1: Performance comparison of the proposed model and baseline models.

Model	ACC	F1	Precision	Recall
Ours	0.8730	0.8743	0.8738	0.8749
LSTM	0.8701	0.8713	0.8700	0.8726
GRU	0.8684	0.8697	0.8688	0.8707
Transformer	0.8452	0.8473	0.8436	0.8532
CNN	0.8444	0.8470	0.8458	0.8484
MLP	0.8419	0.8438	0.8415	0.8469

Compared with LSTM, the proposed model improved accuracy by 0.29 percentage points and macro F1-score by 0.30 percentage points. The difference is modest, but the proposed model remained consistently higher across accuracy, F1-score, precision, and recall. The two recurrent baselines performed better than the standalone CNN, Transformer, and MLP models, which is consistent with the temporal nature of pronunciation, in which the relationship between consecutive mouth configurations carries useful information for classification. The proposed model combines local convolutional processing, bidirectional sequence modeling, and self-attention within a single architecture, and its performance was higher than that of each standalone baseline tested in this study. An ablation study would be needed in future work to determine the individual contribution of each component.

The trained model was further used to score real pronunciation samples. Three examples are shown in Figure 7, Figure 8, and Figure 9. A speaker pronouncing the word “in” obtained a score of 9.9, a speaker pronouncing the word “three” obtained a score of 8.3, and a speaker pronouncing the letter “X” obtained a score of 8.1. These results show that the model can quantify pronunciation quality and provide intuitive feedback to patients during training.

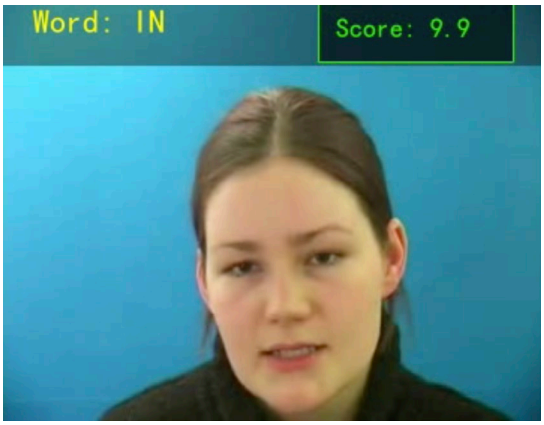


Figure 7: Visual feedback for the word “in” (score 9.9).

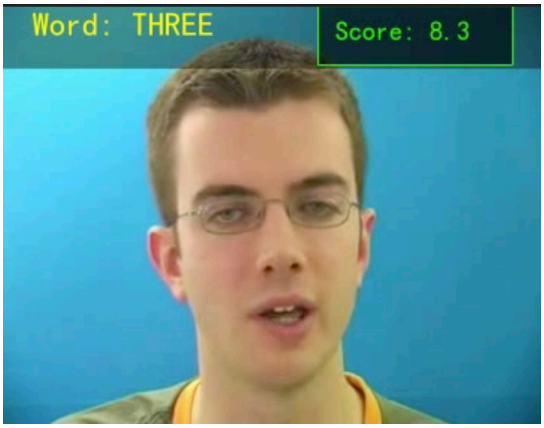


Figure 8: Visual feedback for the word “three” (score 8.3).

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

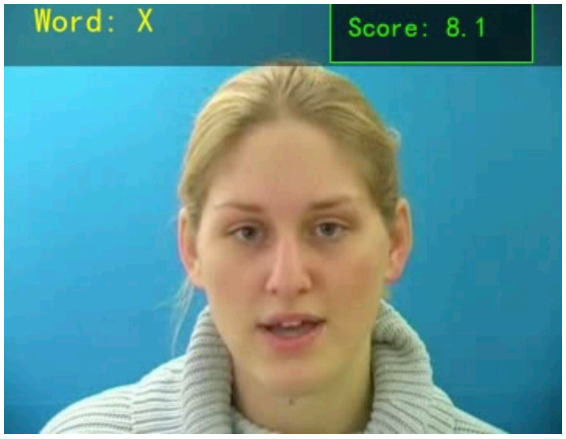


Figure 9: Visual feedback for the letter "X" (score 8.1).

5.2 Hardware Testing

An inflation test was performed to verify the fabrication process of the silicone device. Two molds were used to pour the silicone liquid. One mold contains many small and thin air chambers, and after curing, the two silicone layers were bonded together using liquid silicone. Air was then injected into the device to make it inflate, as shown in Figure 10 and Figure 11. As expected from the design analysis, the first generation could only bend on one surface and could not achieve independent bending of the upper and lower layers.



Figure 10: Inflation test of the pneumatic silicone device (front view).



Figure 11: Inflation test of the pneumatic silicone device (side view).

In the second generation, the line-driven mechanism was introduced to solve this problem. In the first version of the wire-driven design, the Teflon tubing surrounding the fishing lines also bent and deformed, so the device could not bend autonomously. A thicker and stiffer Teflon tube was therefore

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

adopted, enabling autonomous bending of the soft palate section while the area near the teeth remains still. The finished second-generation device is shown in Figure 12, and it successfully achieves compound bidirectional bending controlled by the servo motor.



Figure 12: Second-generation line-driven silicone device.

Two limitations should also be considered. First, GRID consists primarily of adult speakers without cleft palate. The current experiment therefore evaluates the visual analysis pipeline rather than its clinical effectiveness in children with cleft palate, and data from the target population will be necessary before conclusions can be drawn about practical use. Second, the present experiment uses a random 80:20 training-validation split, so samples from the same speaker may appear in both subsets. A speaker-independent evaluation, in which speakers in the evaluation set are completely excluded from training, would provide a stronger test of generalization to unseen individuals.

6. Conclusion

This study presents a soft-palatal training system with an AI-based visual feedback module for children with cleft palate. The hardware part is a line-driven silicone device with pneumatic airbags that can bend bidirectionally to assist soft palate adjustment, and the software part uses facial landmarks to describe mouth geometry and movement. A hybrid 1D-CNN-BiGRU-Transformer architecture classifies four visual pronunciation patterns. On the validation set, the proposed model achieved an accuracy of 87.30%, a macro F1-score of 87.43%, a precision of 87.38%, and a recall of 87.49%, producing the highest values among the LSTM, GRU, Transformer, CNN, and MLP models included in the comparison. The results indicate that the proposed visual feature representation and hybrid temporal model are feasible for analyzing pronunciation-related mouth movements. The current study is an initial technical evaluation. Future work will focus on speaker-independent testing, ablation analysis, data collection from children with cleft palate, and integration of the visual feedback model with the hardware training device.

References

[1] E. B. Stallings, J. L. Isenburg, R. E. Rutkowski, et al., "National Population-Based Estimates for Major Birth Defects, 2016-2020," Birth Defects Research, vol. 116, no. 1, p. e2301, 2024.
[2] A. W. Kummer, "Speech Therapy for Errors Secondary to Cleft Palate and Velopharyngeal Dysfunction," Seminars in Speech and Language, vol. 32, no. 2, pp. 191-198, 2011.
[3] G. Pu and H. Wang, "Review on Research Progress of Machine Lip Reading," The Visual

ISSN 2618-1584 Vol. 8, Issue 5: 41-49, DOI: 10.25236/FMSR.2026.080505

Computer, vol. 39, pp. 3041-3057, 2023.
[4] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in Proceedings of INTERSPEECH, pp. 3652-3656, 2017.
[5] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6548-6552, 2018.
[6] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.
[7] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421-2424, 2006.
[8] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-Time Facial Surface Geometry from Monocular Video on Mobile GPUs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.