Parameter-Efficient Fine-Tuning of Foundation Models for CLP Speech Classification

Susmita Bhattacharjee*, Jagabandhu Mishra†, H.S. Shekhawat* and S. R. Mahadeva Prasanna‡
* Indian Institute of Technology Guwahati, Guwahati
E-mail: sbhattacharjee@iitg.ac.in
† University of Eastern Finland, Joensuu
‡ Indian Institute of Information Technology Dharwad, Dharwad

***Abstract*—We propose the use of parameter-efficient fine-tuning (PEFT) of foundation models for cleft lip and palate (CLP) detection and severity classification. In CLP, nasalization increases with severity due to the abnormal passage between the oral and nasal tracts; this causes oral stops to be replaced by glottal stops and alters formant trajectories and vowel space. Since foundation models are trained for grapheme prediction or long-term quantized representation prediction, they might have the potential to better discriminate CLP severity when fine-tuned on domain-specific data. We conduct experiments on two datasets: English (NMCPC) and Kannada (AIISH). We perform a comparative analysis using embeddings from self-supervised models Wav2Vec2 and WavLM, and the weakly supervised Whisper, each paired with SVM classifiers, and compare their results with traditional handcrafted features—eGeMAPS and ComParE. Finally, we fine-tune the best-performing Whisper model using PEFT techniques: Low-Rank Adapter (LoRA) and Decomposed Low-Rank Adapter (DoRA). Our results demonstrate that the proposed approach for severity classification achieves relative improvements of 26.4% and 63.4% in macro-average F1 score over the best foundation model and handcrafted feature baselines on the NMCPC dataset, and improvements of 6.1% and 52.9% on the AIISH dataset, respectively.**

I. Introduction

Cleft lip and palate (CLP) is a congenital condition affecting the craniofacial region [35], [20], [30]. By 2019, a total of $192,708$ cases of CLP had been reported globally [33]. CLP alters the structure of the oral and nasal cavities, which profoundly impacts speech production [35], [20], [30]. In particular, the opening between the lip and nasal cavity results in speech that is breathy and highly nasalized [36], [34]. In addition, the opening between the oral and nasal cavities shifts the natural resonances of speech [35], [17], making it acoustically distinct from that of normal speakers. To treat this speech disorder effectively, the severity of CLP needs to be accurately detected and graded, which continues to be an active area of research in the medical domain [4].

Researchers aim to detect and grade the severity of CLP directly from speech signals, motivated by the fact that CLP is a speech disorder [31]. In individuals with CLP, the opening between the nasal and oral cavities primarily disrupts articulation and may also alter excitation characteristics [8], [27]. Compared to normal speech, CLP speech often exhibits pressure leakage from the oral to nasal tract during articulation, which leads speakers to replace oral stops such as /p/, /t/, and /k/ with glottal stops and to produce nasalized vowels [17]. Speakers with CLP also tend to produce distorted consonants, weak pressure sounds, and imbalanced resonance [25]. These speech patterns introduce measurable changes in acoustic features, including increased nasal formants, altered formant trajectories, reduced vowel space, and distinctive spectral patterns associated with glottal or pharyngeal substitutions [8], [17].

Although research in this area remains limited, several studies have actively explored the automatic detection and severity classification of CLP speech [32], [29], [10], [9], [24], [1], [21], [15]. Researchers typically follow two main approaches: one applies *signal processing* techniques guided by the articulatory and acoustic characteristics of CLP speech, while the other uses data-driven methods such as *self-supervised or weakly supervised* foundation models to extract high-level representations directly from raw audio for classification tasks. [29] extracted Mel-frequency cepstral coefficients (MFCCs) from variational mode decomposed signals to assess CLP severity. In [10], the authors parameterized the linear prediction residual signal on the vowel /i/ and extracted features such as vocal tract constriction, peak-to-sidelobe ratio, and spectral moments. The study in [1] trained a deep neural network to generate aggregated posteriors of nasalized phonemes. Other works used automatic speech recognition posteriors [32], representations from wav2vec2 models [3], and applied vision transformers [23] to detect and assess the severity of CLP speech. In summary, data-driven methods and methods that utilize representations from foundation models have shown more promising performance compared to traditional signal processing based approaches [3].

Motivated by the promising performance of data-driven approaches, we present a comprehensive analysis of how representations from foundation models specifically, self-supervised (Wav2Vec2 [2], WavLM [5]), and weakly supervised (Whisper [26]) perform in detecting and classifying the severity of CLP speech. While researchers have conducted similar studies in other speech-based biomedical domains such as dementia [12] and dysarthria [6] assessment, no such analysis, to the best of our knowledge, exists for CLP speech. Foundation models are typically trained to predict either *quantized speech units* over long-term dynamics or *speech graphemes* from normal, unimpaired speech. As a result, their learned representations may inherently carry discriminatory features that can

979-8-3315-7206-8/25/$31.00

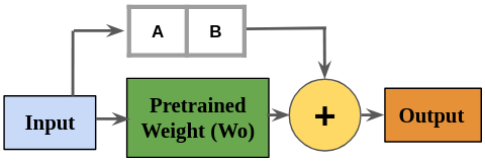

Fig. 1: Block Diagram of LoRA Framework

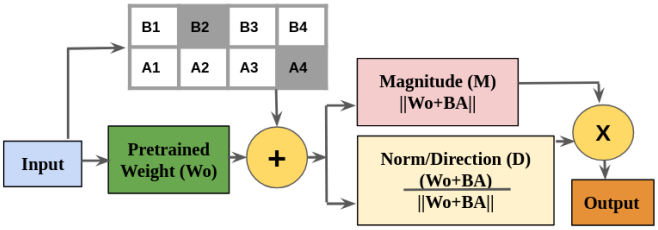

Fig. 2: Block Diagram of DoRA Framework

distinguish CLP speech from normal speech and capture its severity. However, since most foundation models are trained on adult speech [2], [26], [5], and CLP speech data predominantly comes from children [17], this age-domain mismatch, along with the limited availability of CLP-specific data, makes fine-tuning these large models challenging. To address this, we leverage recent advances in *parameter-efficient fine-tuning* (PEFT)[13], which enable effective adaptation to low-resource domains without updating the full model. Specifically, we fine-tune foundation models using *low-rank adaptation (LoRA)* [14] and *weight-decomposed low-rank adaptation (DoRA)* [19] with a classification objective to enhance CLP detection and severity classification performance.

II. Parameter Efficient Finetuning

PEFT refers to a class of techniques that fine-tune large models more efficiently by significantly reducing the number of trainable parameters [13]. Instead of updating the entire network, PEFT methods adapt only specific parts of the model—typically the query, key, and value projection matrices within the transformer's attention blocks [18], [14]. By limiting updates to a small subset of parameters, these methods enable effective adaptation in low-resource settings with limited in-domain data [13]. In this work, we adopt two PEFT methods: LoRA [14] and DoRA [19]. We describe each approach briefly in the following subsections.

A. LoRA (Low-Rank Adaptation)

LoRA [14] is one of the most widely used PEFT methods. It functions as an adapter by introducing trainable parameters only during training, while keeping the model's original size unchanged. Instead of updating the full weight matrix, LoRA decomposes the weight update ΔW into two smaller low-rank matrices, A and B, and learns these through backpropagation. After training, it combines them to approximate the full weight update. Given a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA applies a low-rank decomposition to the weight update such that $\Delta W = BA$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, with $r \ll \min(d, k)$. The model then updates the weight matrix as follows:

$$W = W_0 + BA \tag{1}$$

The block diagram of the weight update procedure using LoRA is depicted in Figure 1.

B. DoRA (Weight-Decomposed Low-Rank Adaptation)

DoRA extends LoRA by decomposing the weight update into two distinct components: magnitude and direction. It refines LoRA's approach by breaking down high-rank LoRA layers into multiple structured single-rank components. During training, DoRA dynamically prunes the less important components, optimizing the parameter budget by retaining only those that contribute meaningfully to the task.Instead of learning a full low-rank update, DoRA uses r pairs of single-rank matrices and continuously evaluates their utility. It removes those with minimal impact, enabling a more efficient and compact adaptation. The updated weight matrix is defined as:

$$W = M \cdot D, \tag{2}$$

where $M = \|W_0 + BA\|$ is a learnable scalar or vector that controls the magnitude of the update, while $D = \frac{W_0+BA}{\|W_0+BA\|}$ is the normalized direction of the low-rank update. Figure 2 illustrates the block diagram of the weight update process using DoRA.

C. Proposed PEFT with foundation models

We are using the foundation model encoder with either LoRA or DoRA adapters. After the final transformer layer, we add a fully connected layer for classification. We apply mean pooling to the output of the last transformer layer to obtain an utterance-level representation, which we feed into the classification layer. During training, we optimize the LoRA or DoRA adapters using cross-entropy loss, while keeping the rest of the encoder frozen. The block diagram of the proposed training procedure is shown in Figure 3.

III. Experimental Setup

We conduct our experiments using two datasets: (1) the New Mexico Cleft Palate Centre (NMCPC) dataset [15], which contains English speech recordings, and (2) the All India Institute of Speech and Hearing (AIISH) dataset [24], which includes Kannada speech recordings. We perform two tasks: *CLP detection* (a binary classification between normal and CLP) and *severity classification* (a four-class classification: normal, mild, moderate, and severe). To establish a baseline, we extract acoustic features using the extended Geneva

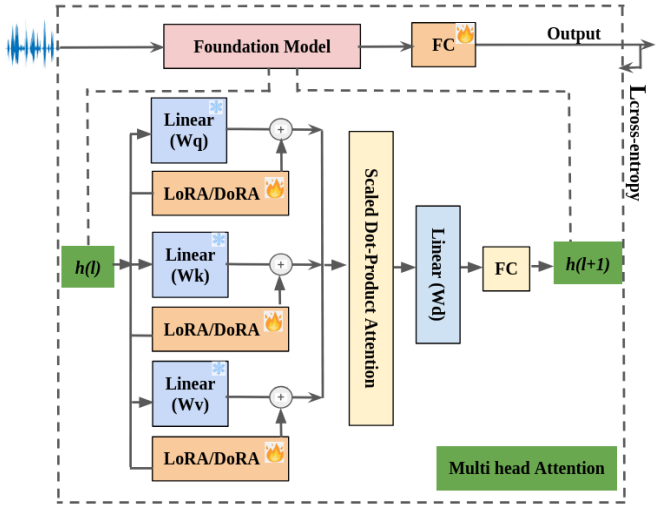

Fig. 3: Proposed framework: Audio input processed by a foundation model, followed by mean pooling and Fully Connected (FC) layers to generate logits with cross-entropy loss. LoRA/DoRA updates linear layers (Wq, Wk, Wv) in multi-head attention while freezing other parameters, enhanced by scaled dot-product attention.

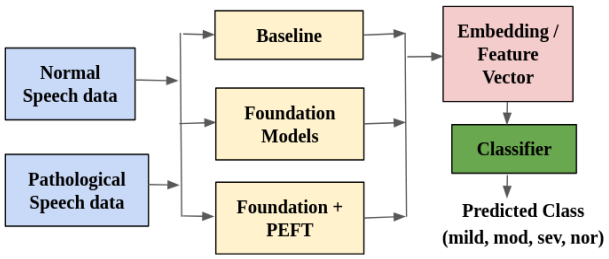

Fig. 4: Block Diagram of the Experiment Design; mod, sev and nor denote moderate, severe and normal respectively.

Minimalistic Acoustic Parameter Set (eGeMAPS) [11] and Computational Paralinguistics Challenge feature set (ComParE) [28] feature sets and train a support vector machine (SVM) classifier [7], inspired by their proven effectiveness in prior biomedical speech studies. Next, we extract pooled representations from the final transformer layer of three foundation models—Whisper[1], Wav2Vec2[2], and WavLM[3]—and use these representations to train SVM classifiers.

Figure 4 illustrates the overall experimental framework. We evaluate model performance using both accuracy and macro-averaged F1 score, giving special importance to the macro-averaged F1 score due to class imbalance. For all experiments, we pool the training and development sets and apply 5-fold cross-validation. We select the model that achieves the highest macro-averaged F1 score for final evaluation on the evaluation set.

[1]https://huggingface.co/openai/whisper-small
[2]https://huggingface.co/facebook/wav2vec2-base-960h
[3]https://huggingface.co/microsoft/wavlm-base

A. Datasets

NMCPC dataset: The New Mexico Cleft Palate Centre collected a speech dataset consisting of 65 speakers, including 41 speakers with CLP—22 male and 19 female—and 24 normal speakers—20 male and 4 female. The dataset includes 76 unique utterances, each containing a maximum of 5 words, and features speakers aged between 9 and 13 years. Clinicians classified the CLP speakers into three severity levels: mild, moderate, and severe [15]. The dataset comprises a total of $1,463$ utterances. The training set contains 929 utterances, including 280 normal, 246 mild, 204 moderate, and 199 severe. The development set includes 235 utterances, with 70 normal, 62 mild, 53 moderate, and 50 severe. The evaluation set comprises 299 utterances, consisting of 89 normal, 77 mild, 67 moderate, and 66 severe. The speakers are disjoint in training, development and evaluation splits.

AIISH dataset: The All India Institute of Speech and Hearing, India, collected a speech dataset comprising 60 speakers, including 31 normal and 29 CLP speakers. Among them, 19 normal female and 12 normal male speakers participated, along with 9 CLP female and 20 CLP male speakers. The dataset includes 19 unique utterances, each containing a maximum of 3 words. All participants were native Kannada speakers between the ages of 7 and 12 years and did not exhibit any other congenital syndromes such as hearing impairment. The dataset contains a total of $2,726$ utterances and is partitioned into training, development, and evaluation sets. The training set consists of $1,731$ utterances, including $1,106$ normal, 302 mild, 247 moderate, and 76 severe utterances. The development set includes 508 utterances, comprising 357 normal, 76 mild, 56 moderate, and 19 severe utterances. The evaluation set contains 487 utterances, with 278 normal, 95 mild, 76 moderate, and 38 severe utterances.

B. Baseline Features and Models

For baseline comparison, we extract two traditional feature sets using openSMILE: eGeMAPS[4], which provides 88 dimensional features [11], and ComParE[4], which yields $6,373$ dimensional features [28]. We also create a merged set that combines both, resulting in a total of $6,461$ dimensional features. To improve computational efficiency, we apply Principal Component Analysis (PCA) on ComParE and the merged set to reduce the handcrafted features to 100 dimensions. We then use these feature sets with the SVM classifier with Radial Basis Function (RBF) [22] kernel to evaluate performance.

C. Foundation Models

We use three state-of-the-art pre-trained models Whisper small [26], Wav2Vec2-base [2], and WavLM-base [5] as frozen feature extractors. For each model, we first resample the raw audio to 16 kHz and then pad or truncate it to a uniform duration of 30 seconds. For Whisper, we convert the audio into log-Mel spectrograms using its built-in AutoProcessor[5].

[4]https://audeering.github.io/opensmile/
[5]https://huggingface.co/docs/transformers/main/en/model_doc/auto

We then extract the last hidden states from the encoder and apply mean pooling to obtain a typically 768-dimensional representation per utterance. We use these pooled representations as input features to an SVM classifier with an RBF kernel for classification.

D. Foundation Models with PEFT

We use the best-performing foundation model augmented with LoRA and DoRA. We introduce the low-rank adapters only in the key, value, and query matrices of the transformer, using a rank of 8. We connect the output of the final transformer encoder to a fully connected layer of size 2 or 4, depending on whether the task is detection or severity classification. We train the model using the AdaM optimizer [16] with a learning rate of $8e{-}5$ and categorical cross-entropy loss.

IV. Results and Discussion

Table I presents the performance of both detection and severity classification using the baseline and proposed methods.

A. CLP Detection

Table I reveals that the eGeMAPS feature set outperforms both ComParE and the merged feature set on both datasets. On the NMCPC dataset, eGeMAPS achieves a classification accuracy of 81.61% and an F1 score of 0.80. In comparison, ComParE achieves only 56.52% accuracy and an F1 score of 0.53, while the merged feature set yields 60.87% accuracy and an F1 score of 0.56. On the AIISH dataset, eGeMAPS again delivers the best performance, achieving 85.63% accuracy and a 0.85 F1 score. ComParE performs lower, with 66.94% accuracy and a 0.64 F1 score, and the merged set slightly improves to 67.56% accuracy and a 0.65 F1 score. These results clearly show that eGeMAPS provides better discrimination between normal and CLP speech than ComParE and its combination with eGeMAPS.

We also observe improved performance with foundation models, where Whisper outperforms both Wav2Vec2 and WavLM. On the NMCPC dataset, Whisper achieves 93.31% accuracy and an F1 score of 0.92, while Wav2Vec2 attains 81.61% accuracy and 0.79 F1 score, and WavLM yields 78.26% accuracy with an F1 score of 0.71. We notice a similar trend on the AIISH dataset. Whisper achieves 93.02% accuracy and a 0.93 F1 score, outperforming Wav2Vec2 with 87.89% accuracy and a 0.87 F1 score, and WavLM with 92.2% accuracy and a 0.92 F1 score. These results indicate that among the foundation models, Whisper provides superior discriminative ability in classifying CLP and normal speech.

Motivated by Whisper's promising performance among foundation models, we evaluated the proposed LoRA and DoRA adaptation methods within the Whisper model. On the NMCPC dataset, Whisper with DoRA achieved a higher accuracy of 94.31% and an F1 score of 0.93, compared to 93.65% accuracy and the same F1 score of 0.93 using LoRA. However, on the AIISH dataset, LoRA outperformed DoRA, achieving 94.25% accuracy and a 0.94 F1 score, while DoRA achieved 93.63% accuracy and a 0.93 F1 score. This reversal in performance between datasets may be attributed to the difference in language—NMCPC contains English speech, while AIISH contains Kannada. Further analysis is necessary in this direction to draw a more detailed and conclusive explanation.

B. Severity Classification

We observe a similar trend in severity classification as in the detection task. However, the performance of severity classification is comparatively lower, likely due to the increased complexity of distinguishing between four classes instead of two. Among the baseline feature sets, eGeMAPS achieves better results, with an F1 score of 0.41 on the NMCPC dataset and 0.34 on the AIISH dataset. In contrast, ComParE yields lower F1 scores of 0.23 and 0.26 on NMCPC and AIISH, respectively, while the merged ComParE and eGeMAPS features achieve 0.23 and 0.28. These results indicate that, similar to the detection task, eGeMAPS offers better discriminative ability for differentiating between severity levels compared to ComParE and the combined feature set.

Using the foundation models, we observe that Whisper embeddings combined with an SVM classifier achieve the best performance in severity classification, similar to the detection task. Whisper attains F1 scores of 0.53 on the NMCPC dataset and 0.49 on the AIISH dataset. In comparison, Wav2Vec2 achieves F1 scores of 0.40 and 0.46, while WavLM yields 0.40 and 0.36 on NMCPC and AIISH, respectively. These results indicate that Whisper embeddings offer superior discriminative ability in differentiating between severity levels compared to Wav2Vec2 and WavLM.

Finally, severity classification follows a similar trend to CLP detection using proposed parameter-efficient fine-tuning with Whisper. On the NMCPC dataset, DoRA achieves a slightly better F1 score of 0.67 compared to 0.65 with LoRA. In contrast, on the AIISH dataset, LoRA performs slightly better, achieving an F1 score of 0.52, while DoRA yields 0.51. These results further support the observation that adaptation methods may behave differently across datasets, possibly due to language variations.

C. Discussions

We began our experiments with baseline features for CLP detection and severity classification. Among these, eGeMAPS combined with an SVM classifier achieved the best performance in terms of both accuracy and F1 score on both the NMCPC and AIISH datasets. We then evaluated foundation models, including the self-supervised Wav2Vec2 and WavLM, as well as the weakly supervised Whisper model. For both detection and severity classification, Whisper embeddings consistently outperformed the others in terms of accuracy and F1 score. Motivated by Whisper's strong performance, we applied our proposed parameter-efficient fine-tuning methods using LoRA and DoRA. We observed that DoRA yielded better results on the NMCPC dataset, while LoRA performed better on the AIISH dataset.

We compared the performance across methods and found that our proposed LoRA and DoRA adaptations applied to

TABLE I: Performance comparison of models across NMCPC and AIISH datasets for CLP detection (normal and clp) and 4-class severity classification task (mild, moderate, severe, and normal). Whisper has a total of 244 million trainable parameters. LoRA and DoRA adapters have about $295k$ and $332k$ trainable parameters, respectively.

	Features	Detection (2 class)				Severity classification (4 class)			
		NMCPC		AIISH		NMCPC		AIISH	
		Accuracy (%)	F1 score	Accuracy (%)	F1 score	Accuracy (%)	F1 score	Accuracy (%)	F1 score
Baseline Features	eGeMAPS	**81.61**	**0.80**	**85.63**	**0.85**	**46.82**	**0.41**	**60.99**	**0.34**
	ComParE	56.52	0.53	66.94	0.64	27.09	0.23	51.95	0.26
	Merged eGeMAPS & ComParE	60.87	0.56	67.56	0.65	25.08	0.23	52.77	0.28
Foundation models	Whisper	**93.31**	**0.92**	**93.02**	**0.93**	**56.86**	**0.53**	**70.02**	**0.49**
	Wav2vec2	81.61	0.79	87.89	0.87	43.81	0.40	67.97	0.46
	WavLM	78.26	0.71	92.20	0.92	41.81	0.40	64.27	0.36
Foundation model+ LoRA/DoRA	Whisper LoRA	93.65	0.93	**94.25**	**0.94**	68.23	0.65	**72.28**	**0.52**
	Whisper DoRA	**94.31**	**0.93**	93.63	0.93	**70.23**	**0.67**	69.82	0.51

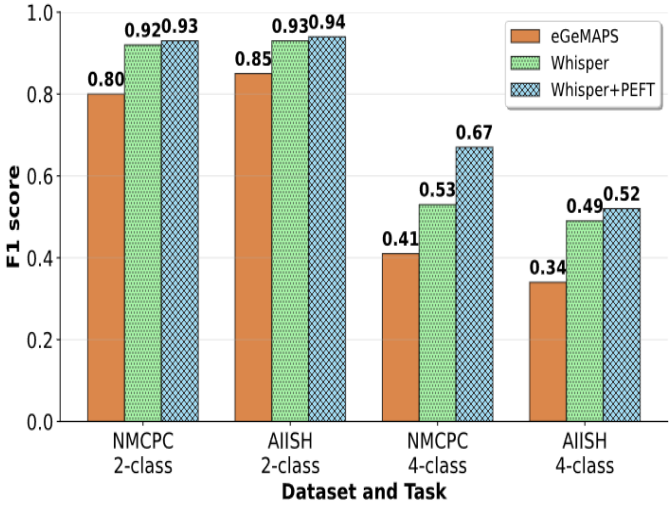

Fig. 5: Performance comparison (F1-scores) of the best-performing models using eGeMAPS, Whisper, and Whisper+PEFT feature sets across the NMCPC and AIISH datasets for CLP detection and severity classification tasks. For AIISH (both tasks), the Whisper+PEFT model uses LoRA, while for NMCPC (both tasks), the Whisper+PEFT model uses DoRA.

the Whisper model outperformed both the pretrained Whisper with SVM and the traditional eGeMAPS features with SVM in CLP detection and severity classification on both the NMCPC and AIISH datasets. On the NMCPC dataset, Whisper with DoRA achieved an F1 score of 0.93 for CLP detection and 0.67 for severity classification. In comparison, the pretrained Whisper with SVM achieved F1 scores of 0.92 and 0.49, while eGeMAPS with SVM reached 0.80 and 0.41. On the AIISH dataset, Whisper with LoRA attained F1 scores of 0.94 for CLP detection and 0.52 for severity classification, compared to 0.93 and 0.49 from Whisper with SVM, and 0.85 and 0.34 from eGeMAPS with SVM. Figure 5 presents a visual comparison of these results. These findings confirm our initial hypothesis that applying LoRA and DoRA for parameter-efficient fine-tuning improves performance in both CLP detection and severity classification over the use of pretrained foundation models and handcrafted feature-based approaches.

V. Conclusions

We conducted a comprehensive study on CLP detection and severity classification using both traditional acoustic features and representations from foundation models. Our experiments on the NMCPC and AIISH datasets showed that eGeMAPS features outperformed ComParE when used with SVM classifiers. Among the foundation models, Whisper consistently achieved higher accuracy and F1 scores than Wav2Vec2 and WavLM. Motivated by Whisper's strong performance, we applied parameter-efficient fine-tuning using LoRA and DoRA adapters. Whisper with DoRA yielded the best results on the English NMCPC dataset, while LoRA performed better on the Kannada AIISH dataset.

These findings highlight the effectiveness of the proposed parameter-efficient fine-tuning strategy for foundation models in both CLP detection and severity classification. In future work, we plan to investigate language-specific fine-tuning to better understand the impact of language characteristics on model performance. We also aim to explore explainability techniques to interpret the model's decisions, particularly in distinguishing between CLP severity levels. These directions will help us develop more transparent, inclusive, and clinically useful models for speech-based health diagnostics.

References

[1] Murni Mohd Amir, Ani Liza Asnawi, Nur Aishah Zainal, and Ahmad Zamani Jusoh. Predicting hypernasality using spectrogram via deep convolutional neural network (dcnn). *2024 IEEE International Conference on Computing (ICOCO)*, pages 398–403, 2024.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[3] Ilja Baumann, Dominik Wagner, Franziska Braun, Sebastian P. Bayerl, Elmar Noth, Korbinian Riedhammer, and Tobias Bocklet. Influence of utterance and speaker characteristics on the classification of children with cleft lip and palate. *INTERSPEECH 2023*, 2022.

[4] Susmita Bhattacharjee and Rohit Sinha. Sensitivity analysis of maskcyclegan based voice conversion for enhancing cleft lip and palate speech recognition. In *2022 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5, July 2022.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518, 2021.

[6] H Christensen, S Cunningham, C Fox, and P Green. Automatic classification of severity of articulation disorders in dysarthric speech. In *Proceedings of Interspeech*, 2012.

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] N Dhananjaya and B Yegnanarayana. Speaker change detection in casual conversations using excitation source features. *Speech communication*, 50(2):153–161, 2008.

[9] Akhilesh Kumar Dubey, S. R. Mahadeva Prasanna, and Samarendra Dandapat. Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features. *The Journal of the Acoustical Society of America*, 146 6:4211, 2019.

[10] Akhilesh Kumar Dubey, Deepak Kumar Singh, and B. B. Tiwari. Hypernasality severity analysis using spectral and residual features. *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, pages 1–6, 2021.

[11] Florian Eyben, Klaus R. Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:190–202, 2016.

[12] Parismita Gogoi, Vishwanath Pratap Singh, Seema Khadirnaikar, Soma Siddhartha, Sishir Kalita, Jagabandhu Mishra, Md Sahidullah, Priyankoo Sarmah, and S. R. M. Prasanna. Leveraging AM and FM rhythm spectrograms for dementia classification and assessment, 2025.

[13] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, 2024, 2024.

[14] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[15] Mohammad Hashim Javid, Krishna Gurugubelli, and Anil Kumar Vuppala. Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6754–6758, 2020.

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

[17] Ann W. Kummer. *Cleft Palate and Craniofacial Anomalies: Effects on Speech and Resonance*. Delmar Cengage Learning, 2007.

[18] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *ArXiv*, abs/2303.15647, 2023.

[19] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *ArXiv*, abs/2402.09353, 2024.

[20] Anette Lohmander and Maria Olsson. Methodology for perceptual assessment of speech in patients with cleft palate: A critical review of the literature. *The Cleft Palate-Craniofacial Journal*, 41:64 – 70, 2004.

[21] Vikram C. Mathad, Nancy J Scherer, Kathy Chapman, Julie M. Liss, and Visar Berisha. A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. *IEEE Transactions on Biomedical Engineering*, 68:2986–2996, 2020.

[22] John Moody and Christian J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[23] Oraphan Nantha, Benjaporn Sathanarugsawait, and Prasong Praneetpolgrang. Enhanced cleft lip and palate classification using siglip 2: A comparative study with vision transformers and siamese networks. *Applied Sciences*, 2025.

[24] K Nikitha, Sishir Kalita, CM Vikram, M Pushpavathi, and SR Mahadeva Prasanna. Hypernasality severity analysis in cleft lip and palate speech using vowel space area. In *Interspeech*, pages 1829–1833, 2017.

[25] & Karnell Peterson-Falzone, Hardin-Jones. *Cleft Palate Speech*. Springer, 2010.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2022.

[27] Mousmita Sarma, Sree Nilendra Gadre, Biswajit Dev Sarma, and SR Mahadeva Prasanna. Speaker change detection using excitation source and vocal tract system information. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[28] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Interspeech*, 2013.

[29] Kilaru Sireesha, Akhilesh Kumar Dubey, D. Govind, Samudravijaya K., and Suryakanth V. Gangashetty. Variational mode decomposition based features for detection of hypernasality in cleft palate speech. *Biomedical Signal Processing and Control*, 97:106689, 2024.

[30] Jackie Stengelhofen. Cleft palate: The nature and remediation of communication problems. *Churchill Livingstone*, 1993.

[31] Protima Nomo Sudro, Rohan Kumar Das, Rohit Sinha, and S. R. Mahadeva Prasanna. Significance of data augmentation for improving cleft lip and palate speech recognition. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 484–490, 2021.

[32] M. VikramC., Ayush Tripathi, Sishir Kalita, and S. R. Mahadeva Prasanna. Estimation of hypernasality scores from cleft lip and palate speech. In *Interspeech*, 2018.

[33] Dawei Wang, Boyu Zhang, Qi Zhang, and Yiping Wu. Global, regional and national burden of orofacial clefts from 1990 to 2019: an analysis of the global burden of disease study 2019. *Annals of Medicine*, 55, 05 2023.

[34] Tara L. Whitehill and Cynthia H F Chau. Single-word intelligibility in speakers with repaired cleft palate. *Clinical Linguistics & Phonetics*, 18:341 – 355, 2004.

[35] D. J. Zajac and L. D. Vallino. *Evaluation and Management of Cleft Lip and Palate: A Developmental Perspective*. Plural Publishing, 2017.

[36] David J. Zajac, Caitrin Plante, Amanda Lloyd, and Katarina L. Haley. Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate. *The Cleft Palate-Craniofacial Journal*, 48:538 – 549, 2011.