Deep Learning-Based Multi-Aspect Pronunciation Assessment for Individuals with Down Syndrome

David Fernández-García, César González-Ferreras, Valentín Cardeñoso-Payo, Mario Corrales-Astorgano

ECA-SIMM Research Group, Universidad de Valladolid, Spain

{david.fernandez, cesargf, valentin.cardenoso, mario.corrales}@uva.es

Abstract

This paper explores the use of an annotated speech corpus to assess multiple dimensions of speech quality—particularly phonetic, fluency and prosody—in individuals with Down syndrome, with the aim of informing the development of automated assessment tools. We conducted a series of experiments using the GOPT model, together with representations extracted from fine-tuning Wav2Vec models focused on phoneme classification. Model predictions were compared against expert annotations from a speech-language pathologist using Pearson correlation. Results demonstrate significant improvements over prior work, with correlations up to 0.49 in certain aspects, particularly for phonetic and fluency dimensions, while prosody remained more challenging to model. The study highlights the potential of Transformer-based architectures for atypical speech assessment and underscores the challenges inherent in assessing atypical speech, particularly due to variability linked to specific disfluency types.

Keywords: pronunciation assessment, Down syndrome, phonetics, fluency, prosody

1. Introduction

Pronunciation assessment plays a fundamental role in educational contexts, particularly in the context of language instruction (Tejedor-García et al., 2020; Eskenazi, 2009). It serves as an essential tool for facilitating the development of learners' spoken language abilities by enabling the provision of targeted, formative feedback. Such feedback is instrumental in helping learners recognize and correct specific pronunciation errors, thereby promoting clearer, more natural, and intelligible speech, which is vital for effective communication and overall language proficiency.

In response to the growing demand for scalable and effective language learning tools, Computer-Assisted Pronunciation Training (CAPT) systems have emerged as a prevailing alternative for second-language (L2) learners, offering objective and personalized feedback in self-directed, stress-free learning scenarios (Van Moere and Downey, 2016). A crucial component of CAPT is Automatic Pronunciation Assessment (APA), which aims to quantify a learner's oral proficiency, often in a "reading-aloud" context where learners pronounce a given text prompt. While early APA efforts typically focused on a single aspect (e.g., accuracy) at a single linguistic level (e.g., phoneme) (Witt and Young, 2000; Coutinho et al., 2016), recent research has shifted towards more comprehensive multi-aspect, multi-granularity models (Gong et al., 2022; Do et al., 2023; Chao et al., 2022). These models simultaneously evaluate diverse aspects (e.g., accuracy, fluency or prosody) across phoneme, word, and utterance levels. A seminal work in this paradigm is the Goodness of Pronunciation Transformer (GOPT) (Gong et al., 2022), a unified neural architecture designed for these complex assessment tasks. The GOPT model is specifically engineered to process phoneme-level Goodness of Pronunciation (GOP) features as its primary input, building upon a foundational metric in the field.

The Goodness of Pronunciation (GOP) metric is a key approach in automated pronunciation assessment, developed to quantify the acoustic deviation of a speaker's utterances from normative speech patterns (Witt and Young, 2000; Hu et al., 2015). While the GOP method is widely employed in the evaluation of non-native L2 speech (Laborde et al., 2016; Ryu and Chung, 2017), a growing body of research has demonstrated its applicability in the assessment of disordered speech (Hair et al., 2021). Some studies have extended the use of GOP-based metrics to the analysis of speech disorders in various clinical populations, including children with apraxia of speech (Shahin et al., 2018; Shahin and Ahmed, 2019), children with cleft lip/palate (Mathad et al., 2021), and individuals with unilateral facial palsy (Pellegrini et al.,

2014). Additionally, enhanced versions of the GOP algorithm have been successfully applied to the evaluation of dysarthric speech (Yeo et al., 2023). These applications underscore the versatility and diagnostic potential of the GOP framework in both second-language and clinical speech assessment contexts. In addition to its relevance in educational contexts, accurate and objective pronunciation assessment is of significant importance for individuals with pathological or disordered speech (Hendrix et al., 2021). Such impairments can substantially hinder speech production and communicative competence, highlighting the need for automated tools to support speech therapy. Speech-language therapists employ various therapeutic approaches to address speech difficulties, many of which can be partially implemented in software-based tools, which may serve as assistive resources during therapy sessions or provide patients with the means to perform supplementary exercises independently.

Individuals with Down Syndrome (DS) often exhibit substantial communication difficulties that can significantly affect their academic performance, social engagement, and overall quality of life. These challenges are multifaceted, arising from a constellation of factors including impaired articulation of speech sounds, atypical prosodic patterns, reduced speech fluency, and distinct voice production characteristics (Kent and Vorperian, 2013).

Speech production in individuals with DS is characterized by a unique combination of delayed developmental patterns and errors not typically observed in neurotypical speech development (Kent et al., 2021). These difficulties affect multiple dimensions of speech, collectively contributing to the markedly reduced speech intelligibility that is characteristic of individuals with DS. From a phonetic standpoint, diminished intelligibility in DS is primarily attributed to articulatory imprecision, motor speech disorders, and phonological deficits (Kumin, 2012). Fluency impairments are also prevalent, with disfluencies such as repetitions, hesitations, and prolongations frequently observed. The incidence of stuttering and cluttering is significantly higher in individuals with DS compared to the general population (Kent et al., 2021). Additionally, prosodic impairments are common, encompassing deficits in rhythm, stress, and intonation. These may include inappropriate pitch variation, monotonic speech, or excessive prosodic fluctuations (Stojanovik, 2011). Given the multidimensional nature of these impairments, a holistic and integrated approach to speech assessment is essential.

The development of automated pronunciation assessment tools for atypical speech is critically dependent on the availability and quality of specialized speech corpora, along with the precision of their annotations. However, such resources are often scarce, posing a significant barrier to progress. Recent advances in Self-Supervised Learning (SSL), especially in conjunction with Transformer-based architectures, have demonstrated high accuracy in the detection of voice disorders, even under data-limited conditions (Bengesi and El-Sayed, 2025). The effectiveness of models such as Wav2Vec (Baevski et al., 2020) and other SSL representations (Chen et al., 2022; Hsu et al., 2021) in data-scarce contexts (Peng et al., 2021, 2023) is particularly significant for DS speech, where annotated corpora are limited. SSL enables models to learn rich representations from vast amounts of unlabeled speech data, which can then be fine-tuned with smaller, labeled pathological datasets. This paradigm shift could significantly accelerate the development of automated assessment tools. Moreover, Wav2Vec is increasingly being explored as a feature extractor in the assessment of other speech disorders, with studies reporting strong correlations with speech quality assessment (Nguyen et al., 2024).

In this work, we present an unified automated pronunciation assessment model capable of simultaneously evaluating multiple aspects of pronunciation, specifically designed for individuals with DS. It focuses on the application of DNNs to evaluate phonetic, fluency, and prosodic dimensions of speech. The performance of the proposed approach is validated against expert human evaluations.

The structure of the paper is as follows. We begin describing the methodology in section 2, including a description of the dataset, the model, the audio representations and the evaluation metrics. Following this, we present the results of the experiments in section 3. Finally, we conclude with a discussion (section 4) and conclusions and suggestions for future work (section 5).

2. Methodology

2.1. Data

To carry out this study, we used the **PRAUTOCAL** corpus (Escudero-Mancebo et al., 2022). This corpus contains 120 minutes of audio recordings of individuals with Down syndrome in Spanish, distributed across a total of 2,048 different audio files. The recordings were obtained from interactions between people with DS and a video game. The game consisted of 40 different activities, each requiring a fixed spoken response (i.e., a specific sentence to be pronounced).

All audio content was evaluated and annotated

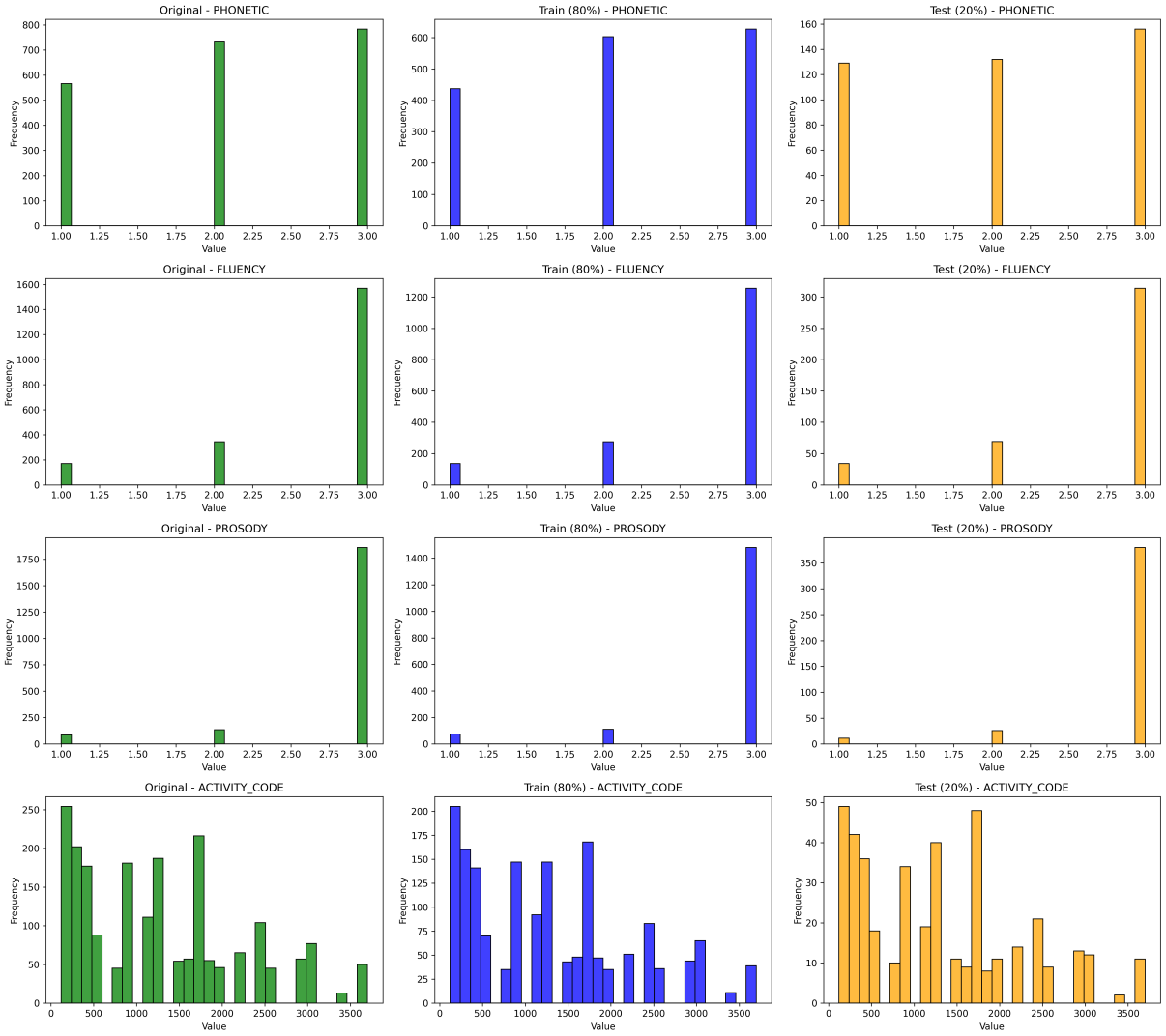

Figure 1: Class distribution by partition.

by a professional linguist, focusing on various aspects of phonetics, fluency, and prosody. The linguist listened to each audio clip and annotated it using a predefined rubric. A detailed description of this rubric can be found in (Corrales-Astorgano et al., 2024). This experiment focuses on 11 aspects, which are grouped into three categories:

- **General aspects**: These include overall scores for each dimension, **phonetics** (PHO), **fluency** (FLU), and **prosody** (PRO). The scoring is categorical and ranges from 1 to 3. Where 3 is best score and 1 is worst.
- **Fluency aspects**: These include specific indicators of fluency such as **blocks** (BLO), **prolongations** (PLG), **repeated sounds** (REPs), **repeated words** (REPw), and **interjections** (INT). Each of these was rated on a scale from 0 to 2, where 0 indicates absence, 1 indicates one occurrence, and 2 indicates two or more occurrences.
- **Prosody aspects**: These include specific prosodic features such as **stress** (STR), **modality** (MOD), and **phrasing** (PHR). Like fluency aspects, they are rated categorically from 0 to 2.

In accordance with established methodologies (Muraina, 2022; Haque et al., 2024), the dataset was partitioned into training and testing sets using a conventional 80/20 ratio. A crucial factor for achieving good performance is class balance in both partitions.

Since it is not feasible to balance all 11 aspects simultaneously, we chose to balance the partitions based only on the **general aspects** and **activities** (see Figure 1). This choice is justified by the assumption that general aspects broadly reflect the more specific features. Thus, balancing the data based on these general values should help ensure overall balance. With regard to the activities, it is also essential to maintain phoneme balance across both partitions. Otherwise, the model could

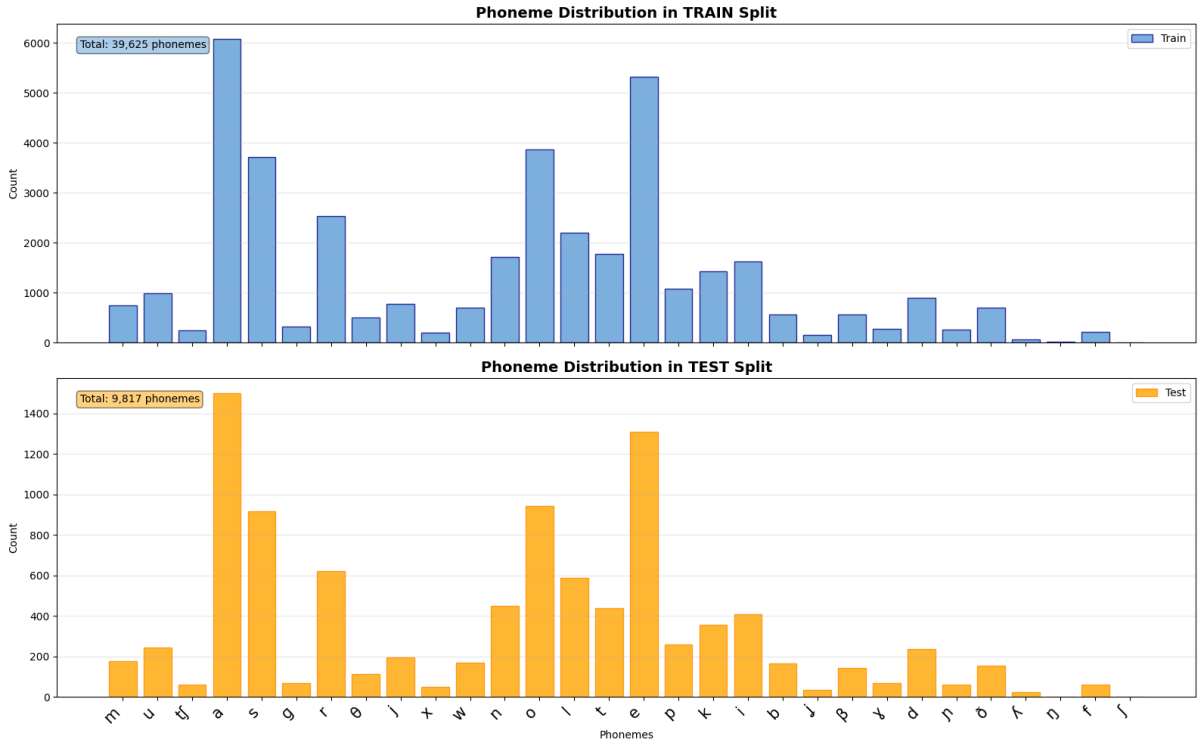

Figure 2: Phoneme distribution for each partition.

overfit to a non-representative phoneme distribution. Since each activity corresponds to a unique sentence, balancing the number of samples from each activity across the sets ensures a phonetically balanced dataset (see Figure 2).

Finally, it is worth noting that the phonetic segmentation of the corpus was obtained using the Montreal Forced Aligner (McAuliffe et al., 2017) as an automatic segmentation tool.

2.2. Model

The architecture used in this study is the **GOPT** model (Gong et al., 2022) (see Figure 3). This architecture links the vector representations of each phoneme in a sentence via an encoder based on attention mechanisms, enabling evaluation of the sentence across multiple speech aspects. The original work supports evaluation at the phrase, word, and phoneme levels. However, in this work all the annotations are at the phrase level, so we only perform global sentence-level evaluation. Evaluations are conducted using regression heads applied to Classify Tokens (CLS). In our case, the model has 11 regression heads and CLS tokens, one for each aspect to be assessed. The phoneme-level representations from the encoder are discarded since no annotations exist at that level.

A hyperparameter tuning process over the train set determined the following optimal values: *epochs*: 15, *learning rate*: 0.001, *batch size*: 25, *number of attention heads*: 1, *head depth*: 3, *encoder input dimension*: 24, and *MSE* as loss function. To mitigate execution variability, we followed the original paper's recommendation of averaging results across 5 runs.

2.3. Audio representations

To apply the GOPT model and paradigm to the PRAUTOCAL dataset, we need phoneme-level vector representations. We adopted two approaches inspired by Yeo et al. (2023) (see Figure 3). That study fine-tunes a Wav2Vec model for phoneme recognition by extracting the feature convolutional encoder and adding an MLP layer with an output size equal to the number of phoneme classes. Fine-tuning is performed on the Common Phone corpus (Klumpp et al., 2022), and the resulting phoneme probabilities are used to compute several variants of the *Goodness of Pronunciation (GoP)* metric (Witt and Young, 2000), enabling sentence-level evaluation. A similar approach applied to the PRAUTOCAL corpus was described in Corrales-Astorgano et al. (2024).

We adopted a similar strategy to generate phoneme-specific representations using a classifier but approached it from two perspectives. The first approach is similar to the one described above: we performed two fine-tunings of the **Wav2Vec-XLRS-53** model (Ruder et al., 2019), each using

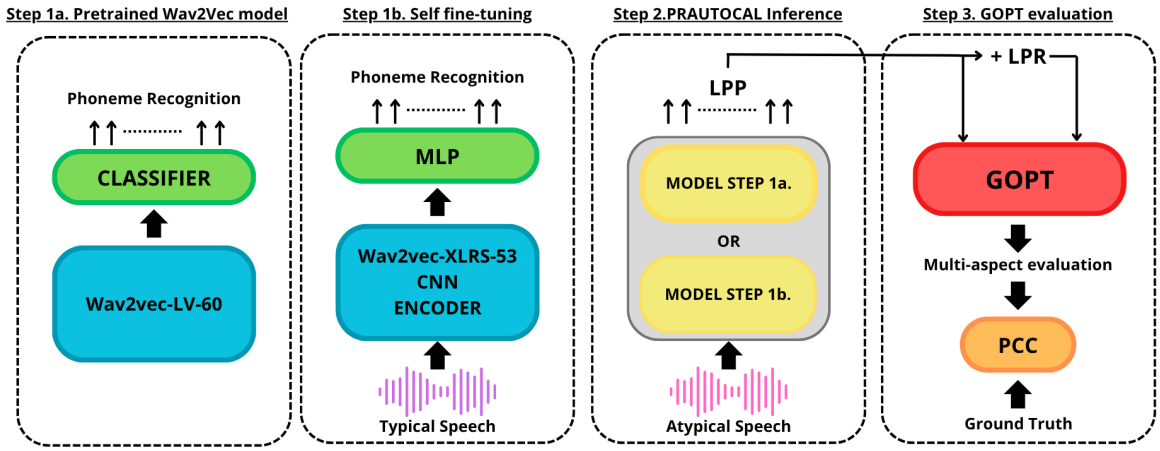

Figure 3: *Overview of the experimental workflow followed in this study. PCC means Pearson Correlation Coefficient*

a different loss function. The choice of this model was motivated by two key factors: the multilingual nature of the Common Phone corpus and the model's demonstrated success in analyzing disordered speech (Hernandez et al., 2022; Leivaditi et al., 2024; Hu et al., 2024). Since many phonemes in Common Phone are not found in PRAUTOCAL, whose phonetic dictionary contains just 30 phonemes, we removed all non-matching phonemes and the silence phoneme, training the model only on phonemes in the PRAUTOCAL dictionary. This avoids unnecessary task complexity and results in 30 output probabilities.

Thus, we trained the model on a filtered variant of the Common Phone corpus, consisting of 2,564,839 labeled phonemes, using the following hyperparameters: *epochs*: 1, *learning rate*: 0.001, *batch size*: 8.

We experimented with two loss functions: **Negative Log-Likelihood Loss** (NLL), which is the one used by default in the reference work, and **Focal Loss** (Lin et al., 2017). Focal Loss was chosen because NLL tends to produce extreme probability values (i.e., a single phoneme with probability 1 and all others with 0), which is not ideal for GOPT, as the model benefits from probabilistic differences among phonemes rather than hard classifications.

The second approach uses a phoneme classification model pretrained with extensive audio data. This is **Wav2Vec2-lv-60-espeak-cv-ft**, which has been self-pretrained on *LibriSpeech* (Panayotov et al., 2015), and then fine-tuned for phoneme recognition task with *Common Voice* (Ardila et al., 2020).

Once the final model is obtained by either method, we run inference on PRAUTOCAL to obtain phoneme-level representations. Following the method described in Gong et al. (2022), we use only the model logits (not the probabilities), which constitute the Log Posterior Probability (**LPP**) representation:

$$LPP(p_k) \approx \frac{1}{1+(t_e-t_s)} \sum_{t=t_s}^{t_e} \log p(p_k \mid o_t) \quad (1)$$

Following the same paper, we augment this representation by concatenating it with the Log Posterior Ratio (**LPR**), which subtracts the known phoneme's logit from the others. The final concatenated vector, known as **GoP Features** (Wang and Lee, 2012).

$$LPR(p_j \mid p_i) = \log p(p_j \mid \mathbf{o}; t_s, t_e) - \log p(p_i \mid \mathbf{o}; t_s, t_e) \quad (2)$$

2.4. Evaluation metrics

The systems described above are evaluated calculating the *Pearson Correlation Coefficient* between the model predictions and the previously defined ground truth. This metric is appropriate for our task, since the goal is to emulate the categorical evaluations of a professional linguist. Pearson's coefficient provides an objective measure of similarity between our model's predictions and expert judgments.

Furthermore, this metric is widely used in speech quality assessment research (Chao et al., 2022; Kadam et al., 2022; Liu et al., 2025). As also noted in Gong et al. (2022), it performs well with unbalanced datasets, which applies to our scenario.

3. Results

Our experiments include three different setups, involving two models and two loss functions. The results are summarized in Table 1, which breaks

MODEL		GENERAL			FLUENCY					PROSODY			MEAN
		PHO	FLU	PRO	BLO	PLG	REPs	REPw	INT	STR	PHR	MOD	
(Corrales-Astorgano et al., 2024)	GoP	0.198	-	-	-	-	-	-	-	-	-	-	-
W2V-XLRS-53-NLL	LPP	0.414 ±0.028	0.465 ±0.026	0.246 ±0.049	0.453 ±0.013	**0.246** ±0.028	0.354 ±0.028	0.373 ±0.018	0.098 ±0.043	-0.003 ±0.007	0.293 ±0.055	**0.133** ±0.033	**0.279** ±0.030
	GoP Feat.	0.416 ±0.028	0.469 ±0.019	**0.275** ±0.021	0.466 ±0.014	0.216 ±0.016	**0.389** ±0.015	0.342 ±0.024	0.066 ±0.037	0.003 ±0.015	0.281 ±0.029	0.106 ±0.039	0.275 ±0.023
W2V-XLRS-53-FOCAL	LPP	**0.425** ±0.043	0.450 ±0.031	0.262 ±0.024	0.441 ±0.021	0.215 ±0.020	0.355 ±0.042	0.369 ±0.018	0.080 ±0.036	-0.017 ±0.031	0.303 ±0.042	0.083 ±0.065	0.269 ±0.034
	GoP Feat.	0.419 ±0.038	0.443 ±0.022	0.201 ±0.023	0.461 ±0.015	0.226 ±0.025	0.360 ±0.023	0.341 ±0.042	0.110 ±0.035	-0.016 ±0.017	0.233 ±0.067	0.065 ±0.013	0.258 ±0.029
W2V-LV-60-espeak-CV-FT	LPP	0.314 ±0.042	**0.493** ±0.013	0.252 ±0.036	**0.467** ±0.009	0.243 ±0.029	0.336 ±0.028	**0.383** ±0.016	**0.153** ±0.015	**0.018** ±0.027	0.294 ±0.019	0.082 ±0.034	0.275 ±0.024
	GoP Feat.	0.297 ±0.042	0.457 ±0.027	0.273 ±0.071	0.436 ±0.037	0.210 ±0.029	0.332 ±0.034	0.363 ±0.023	0.110 ±0.025	-0.036 ±0.035	**0.355** ±0.041	0.097 ±0.034	0.263 ±0.036

Table 1: Pearson Correlation Coefficient (PCC) results for the GOPT-based pronunciation assessment models across 11 evaluated aspects, grouped into General, Fluency, and Prosody categories. Each model was tested using two types of phoneme-level representations: LPP (Log Posterior Probability) and GoP Features (LPP+LPR).

down the results by model and representation, showing performance with LPP only and with the full set of *GoP Features*.

First of all, we can see how our approach doubles the performance of the previous work (Corrales-Astorgano et al., 2024). It should be noted that the previous work focused only on phonetic general scores, whereas the present study evaluates 10 additional aspects.

Overall, we observe significant variability in correlation values across the different aspects. In some cases, we achieve correlations close to 0.500, while others yield near-zero or even negative correlations. The aspects most effectively assessed by the model, regardless of representation type, are **Phonetic**, **Fluency**, and **Blocks**, with correlations exceeding 40% in all three experiments. Conversely, the most challenging aspects to evaluate are **Accent**, **Interjections**, and **Modulation**, two of which belong to the **Prosody** category—the least correlated among the general assessments.

Differences between models are minimal, with a maximum average variation of 0.021 between the best and worst-performing models, as compared to an average standard deviation of 0.03 across the 5 runs for all cases. The top model yields a mean correlation of 0.279, obtained using the LPP representation from our custom fine-tuned *Wav2Vec-XLRS-53* with *Negative Log-Likelihood Loss*. The lowest correlation, 0.258, results from using *Focal Loss* instead. This suggests that our hypothesis about improving results through loss function variation was not validated.

Finally, we note that using *GoP Features* did not yield better performance in this application, contrary to what has been reported in other studies (Do et al., 2023; Gong et al., 2022; Chao et al., 2022). In all three models, performance with LPP alone surpassed that obtained with the full *GoP Feature* set.

4. Discussion

The results underscore the inherent difficulty of the task, as we achieved only moderate to weak correlations even in the best cases. Nevertheless, our findings represent a clear improvement over previous work on the same dataset (Corrales-Astorgano et al., 2024). Our work also extends the evaluation to 10 additional aspects, whereas the prior study was limited to *General Phonetic Aspect*. That study followed a similar procedure, fine-tuning a model for phoneme recognition to generate individual phoneme representations and subsequently applying techniques based on Goodness of Pronunciation (GoP) metrics for evaluation. The performance improvement achieved in this work—an increase in Pearson Correlation Coefficient (PCC) from 0.198 to 0.425 for *General Phonetic Aspect*—demonstrates the effectiveness of the Transformer architecture and its attention mechanisms (Vaswani et al., 2017). These models continue to outperform their predecessors in such tasks (Lee et al., 2024; Yan et al., 2024). Furthermore, our results indicate that the system benefits from the multi-aspect evaluation, a finding consistent with previous literature (Gong et al., 2022; Do et al., 2023; Chao et al., 2022).

Our models exhibit variable performance across the different assessed aspects, with some showing a significantly lower correlation with expert predictions than others. This behavior is consistent with previous work, such as (Gong et al., 2022) on the GOPT model, which encountered a similar challenge with metrics like *Word Stress* and *Utterance Completeness* when operating on the SpeechOcean762 corpus (Zhang et al., 2021). The au-

thors attributed this low correlation to the unbalanced distribution of scores for these aspects in the training data.

An analogous pattern emerges in our results. The *General Prosody Aspect* yielded the lowest Pearson Correlation Coefficient (PCC) of 0.275, in contrast to the 0.425 and 0.493 achieved by Phonetics and Fluency, respectively. As illustrated in Figure 1, prosody is the aspect with the most unbalanced distribution, which supports the data impact hypothesis. However, the fact that Fluency—whose score distribution is also unbalanced and more similar to that of prosody than to phonetics—achieves the highest correlation suggests that data distribution is not the sole determining factor. Other causes could include the inherent difficulty of modeling certain abstract aspects, challenges in optimizing the loss function, or the need to enrich the input representations with more specific information, as proposed in other studies (Chao et al., 2022).

The three models produced highly similar results, indicating that the specific architectural and training modifications had a negligible effect on overall performance. The maximum observed variance in mean values was only 0.021. We attribute this lack of significant differentiation to the fact that all three approaches are built upon the same core principle of phoneme classification. Since the underlying experimental design was consistent and alterations to the models and loss function did not lead to notable performance gaps, this uniformity in results is logical.

Contrary to results from similar studies (Gong et al., 2022; Do et al., 2023), our findings indicate that the LPP representation consistently surpasses GoP Features across all our experiments. We attribute the suboptimal performance of the GoP Features to the low discriminative capacity of the LPPs for this task, whose classification accuracy, with any representation, was limited to under 40%. The LPR calculation, which relies on these LPPs, likely introduces significant noise by making faulty comparisons, thus degrading the feature's effectiveness. This outcome is unsurprising given the inherent difficulty of analyzing speech from individuals with Down Syndrome, which is substantially more complex and variable than the L2 speech examined in the referenced papers.

5. Conclusions and future work

Our results confirm that while evaluating the speech of individuals with Down syndrome is a highly challenging task, Transformer-based models demonstrate considerable potential and a significant improvement over previous approaches, also benefiting from a multi-aspect evaluation framework. A key contribution of this work is the finding that, contrary to studies on second language (L2) speech, the LPP representation consistently outperforms GoP Features. We attribute this result to the high complexity and heterogeneity of speech associated with Down syndrome, which appears to degrade the performance of GoP metrics by introducing noise into their components. This underscores that effective representations for typical or L2 speech could not be directly transferable to atypical speech, reinforcing the critical importance of exploring and developing more robust and adapted representations.

Future work should prioritize the development of new feature representations that are robust to atypical speech and surpass traditional metrics, alongside increasing the quantity and quality of annotated pathological data. Additionally, techniques to mitigate data imbalance, such as data augmentation, should be explored to improve performance on complex aspects like prosody. It will also be crucial to conduct detailed ablation studies and qualitative analyses to identify the most difficult features to model and to explore representations that do not rely solely on phoneme classification in order to capture the supra-segmental information essential for a holistic speech assessment.

6. Acknowledgements

This work was carried out in the Project PID2021-126315OB-I00 that was supported by MCIN / AEI / 10.13039/501100011033 / FEDER, EU.

7. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*, pages 4218–4222.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Staphord Bengesi and Hoda El-Sayed. 2025. Voice disorder prediction with convolutional neural network (cnn). In *Computational Science and Computational Intelligence*, pages 104–112. Springer Nature Switzerland.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2022. 3M: An effective multi-view, multi-granularity, and multi-

aspect modeling approach to English pronunciation assessment. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 575–582. IEEE.

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, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Mario Corrales-Astorgano, César González-Ferreras, David Escudero-Mancebo, Lourdes Aguilar, Valle Flores-Lucas, Valentın Cardenoso-Payo, and Carlos Vivaracho-Pascual. 2024. Pronunciation assessment and automated analysis of speech in individuals with Down syndrome: Phonetic and fluency dimensions. In *IberSPEECH*, pages 26–30.

Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native speakers of english: Measures and feature sets.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Hierarchical pronunciation assessment with multi-aspect attention. In *ICASSP*, pages 1–5.

David Escudero-Mancebo, Mario Corrales-Astorgano, Valentín Cardeñoso-Payo, Lourdes Aguilar, César González-Ferreras, Pastora Martínez-Castilla, and Valle Flores-Lucas. 2022. Prautocal corpus: a corpus for the study of Down syndrome prosodic aspects. *Language Resources and Evaluation*, 56(1):191–224.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844. Spoken Language Technology for Education.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *ICASSP*, pages 7262–7266.

Adam Hair, Guanlong Zhao, Beena Ahmed, Kirrie J Ballard, and Ricardo Gutierrez-Osuna. 2021. Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions. In *Interspeech*, pages 2936–2940.

Alimul Haque, Shams Raza, Sultan Ahmad, Alamgir Hossain, Hikmat AM Abdeljaber, AEM Eljialy, Sultan Alanazi, and Jabeen Nazeer. 2024. Implication of different data split ratio on the performance of model in price prediction of used vehicles using regression analysis. *Data Metadata*, 3:425.

James A Hendrix, Angelika Amon, Leonard Abbeduto, Stamatis Agiovlasitis, Tarek Alsaied, Heather A Anderson, Lisa J Bain, Nicole Baumer, Anita Bhattacharyya, Dusan Bogunovic, et al. 2021. Opportunities, barriers, and recommendations in down syndrome research. *Translational science of rare diseases*, 5(3-4):99–129.

Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Noeth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. 2022. Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. In *Interspeech 2022*, pages 51–55.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, et al. 2024. Self-supervised asr models and features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3561–3575.

Wenping Hu, Yao Qian, Frank K Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.

Sandeep U. Kadam, Anand katri, Vajid N Khan, Avaneesh Singh, Dattatray G. Takale, and Dattatray S. Galhe. 2022. Improve the performance of non-intrusive speech quality assessment using machine learning algorithms. *NeuroQuantology*, 20(10):12937–12944.

Ray D. Kent and Houri K. Vorperian. 2013. Speech impairment in Down syndrome: A review. *Journal of Speech, Language, and Hearing Research*, 56(1):178–210.

Raymond D. Kent, Julie Eichhorn, Erin M. Wilson, Youmi Suk, Daniel M. Bolt, and Houri K. Vorperian. 2021. Auditory-perceptual features of speech in children and adults with Down syndrome: A speech profile analysis. *Journal*

of Speech, Language, and Hearing Research, 64(4):1157–1175.

Philipp Klumpp, Tomas Arias, Paula Andrea Pérez-Toro, Elmar Noeth, and Juan Orozco-Arroyave. 2022. Common phone: A multilingual dataset for robust acoustic modelling. In *LREC*, pages 763–768.

Libby Kumin. 2012. *Early communication skills for children with Down syndrome: A guide for parents and professionals*, 3rd edition edition. Woodbine House.

Vincent Laborde, Thomas Pellegrini, Lionel Fontan, Julie Mauclair, Halima Sahraoui, and Jérôme Farinas. 2016. Pronunciation assessment of japanese learners of french with gop scores and phonetic information. In *Annual conference Interspeech (INTERSPEECH 2016)*, pages 2686–2690.

Haeyoung Lee, Sunhee Kim, and Minhwa Chung. 2024. Analysis of various self-supervised learning models for automatic pronunciation assessment. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Spyretta Leivaditi, Tatsunari Matsushima, Matt Coler, Shekhar Nayak, and Vass Verkhodanova. 2024. Fine-tuning strategies for dutch dysarthric speech recognition: Evaluating the impact of healthy, disease-specific, and speaker-specific data. In *Interspeech 2024*, pages 1295–1299. ISCA.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2999–3007.

Miao Liu, Jing Wang, Fei Wang, Fei Xiang, and Jingdong Chen. 2025. Non-intrusive speech quality assessment based on deep neural networks for speech communication. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):174–187.

Vikram C Mathad, Tristan J Mahr, Nancy Scherer, Kathy Chapman, Katherine C Hustad, Julie Liss, and Visar Berisha. 2021. The impact of forced-alignment errors on automatic pronunciation evaluation. In *Interspeech*, pages 1922–1926.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, pages 498–502.

Ismail Muraina. 2022. Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In *7th international Mardin Artuklu scientific research conference*, pages 496–504.

Tuan Nguyen, Corinne Fredouille, Alain Ghio, Mathieu Balaguer, and Virginie Woisard. 2024. Exploring asr-based wav2vec2 for automated speech disorder assessment: Insights and analysis. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 975–982.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210.

Thomas Pellegrini, Lionel Fontan, Julie Mauclair, Jérôme Farinas, and Marina Robert. 2014. The goodness of pronunciation algorithm applied to disordered speech. In *Interspeech*, pages 1463–1467.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhang. 2021. A study on fine-tuning wav2vec2. 0 model for the task of mispronunciation detection and diagnosis. In *Interspeech*, volume 2021, pages 4448–4452.

Linkai Peng, Yingming Gao, Rian Bao, Ya Li, and Jinsong Zhang. 2023. End-to-end mispronunciation detection and diagnosis using transfer learning. *Applied Sciences*, 13(11):6793.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Hyuksu Ryu and Minhwa Chung. 2017. Mispronunciation diagnosis of l2 english at articulatory level using articulatory goodness-of-pronunciation features. In *SLaTE*, pages 65–70.

Mostafa Shahin and Beena Ahmed. 2019. Anomaly detection based pronunciation verification approach using speech attribute features. *Speech Communication*, 111:29–43.

Mostafa Ali Shahin, Beena Ahmed, Jim X Ji, and Kirrie J Ballard. 2018. Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features. In *Interspeech*, pages 1671–1675.

Vesna Stojanovik. 2011. Prosodic deficits in children with Down syndrome. *Journal of Neurolinguistics*, 24(2):145–155.

Cristian Tejedor-García, David Escudero-Mancebo, Enrique Cámara-Arenas, César González-Ferreras, and Valentín Cardeñoso-Payo. 2020. Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies*, 13(2):269–282.

Alistair Van Moere and Ryan Downey. 2016. 21. technology and artificial intelligence in language assessment. *Handbook of second language assessment*, pages 341–358.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yow-Bang Wang and Lin-Shan Lee. 2012. Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In *ICASSP*, pages 5049–5052.

S.M Witt and S.J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2):95–108.

Bi-Cheng Yan, Jiun-Ting Li, Yi-Cheng Wang, Hsin Wei Wang, Tien-Hong Lo, Yung-Chang Hsu, Wei-Cheng Chao, and Berlin Chen. 2024. An effective pronunciation assessment approach leveraging hierarchical transformers and pre-training strategies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1737–1747, Bangkok, Thailand. Association for Computational Linguistics.

Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. 2023. Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification. In *Interspeech*, pages 166–170.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Interspeech 2021*, pages 3710–3714.