

# Assessing Dysarthric Speech Intelligibility via Generative Augmentation and Self-Supervised Linguistic Representations

Davide Lillini<sup>1</sup>, Carlo Aironi<sup>1</sup>, Lucia Migliorelli<sup>2</sup>, Leonardo Gabrielli<sup>1</sup>, and Stefano Squartini<sup>1</sup>

<sup>1</sup> Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

`d.lillini@pm.univpm.it`

`{c.aironi,l.gabrielli,s.squartini}@staff.univpm.it`

<sup>2</sup> Department of Political Science, Università degli Studi di Teramo, Teramo, Italy

`lmigliorelli@unite.it`

**Abstract.** The development of robust automated systems for assessing speech intelligibility levels in patients with dysarthria is currently hindered by the scarcity of labeled pathological datasets. This study proposes a novel framework addressing this limitation by integrating self-supervised learning with generative data augmentation. We employ Wav2Vec 2.0, a pre-trained self-supervised speech model, to extract and analyze contextualized representations, aligned with both grapheme and phoneme targets, systematically identifying which linguistic units serve as the most effective biomarkers for binary intelligibility classification. To address data scarcity, we employ StyleTTS-2, a diffusion-based generative model, to generate synthetic dysarthric speech samples conditioned on reference audio from the TORGO dataset and text prompts from the UA-Speech dataset. Experimental results demonstrate that contextualized representations yield superior classification accuracy compared to approaches operating on full word- or utterance-level audio. Significantly higher performance is observed on units most degraded by dysarthric impairment, validating the proposed fine-grained approach. Furthermore, synthetic data augmentation proved to enhance model robustness, effectively mitigating data scarcity issues.

**Keywords:** Dysarthria assessment · Speech Intelligibility · Generative Speech Augmentation · Self-Supervised Representation · Text-to-Speech algorithms.

## 1 Introduction

Dysarthria is a motor speech disorder caused by impaired neuromuscular control, affecting multiple speech subsystems and leading to reduced intelligibility, with an impact on patients' quality of life. Speech intelligibility is traditionally assessed through perceptual evaluations by speech-language pathologists;

however, these assessments may be subjective and limited in reproducibility, motivating the development of clinical decision support systems [30,7].

Advances in Digital Signal Processing (DSP) and Machine Learning (ML) enabled the extraction of acoustic biomarkers for assessing speech impairment severity. However, the development of robust speech intelligibility classification models – closely linked to dysarthria severity – is hindered by the scarcity of high-quality labeled pathological datasets [8,11,18].

Different approaches have been proposed in the literature to address intelligibility assessment. Early methods relied on Automatic Speech Recognition (ASR) metrics as proxies for intelligibility, exploiting correlations between word recognition rates and expert ratings [20,27]. More recent studies adopted Deep Learning (DL) architectures, including Convolutional and Recurrent Neural Networks, often combined with transfer learning from ASR models trained on non-pathological speech [12,29]. However, the performance of these models is often overestimated due to speaker leakage between training and testing sets, leading to inflated generalization metrics [18].

Self-Supervised Learning (SSL) approaches, such as Wav2Vec 2.0 [1], provide transferable speech representations from large-scale unlabeled data, but their discriminative role in dysarthric speech intelligibility classification remains underexplored [21,23].

In parallel, generative TTS architectures like StyleTTS-2 [16] enable expressive synthesis and controlled data augmentation. Yet, the impact of synthetic dysarthric speech on model robustness and generalization requires further investigation, particularly regarding optimal augmentation strategies for pathological speech analysis.

Given these premises, this work focuses on two main objectives. We first investigate the internal representations of Wav2Vec 2.0 in order to identify which specific phonetic units encode the most discriminative information for dysarthric speech intelligibility. We distinguish between two types of units based on their orthographic representation: graphemes (individual letters of the Roman alphabet, hereinafter referred to as “graphemes”) and multi-letter phonemes (phonemes whose IPA representation requires multiple letters or special symbols, hereinafter referred to simply as “phonemes”). Representations aligned with graphemes reflect transcription-based intelligibility assessment, while representations aligned with phonemes are analyzed as a proxy for perceptual listening-based evaluation. Second, we evaluate whether the integration of synthetically generated speech data helps improve the robustness and performance of intelligibility classification models.

The remainder of this paper is organized as follows. Section 2 describes the datasets and the proposed methodology. Section 3 outlines the experimental protocol. Section 4 presents and discusses the results. Section 5 concludes the paper and outlines future research directions.

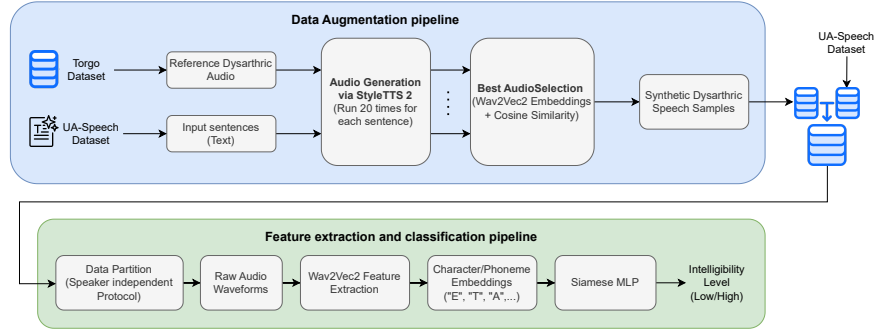


**Fig. 1.** Overview of the proposed architecture, composed of two primary stages. The Data Augmentation pipeline (top, blue background) employs style- and text-conditioned speech generation to produce synthetic dysarthric samples; text prompts are generated by the Gemini Pro LLM conditioned on the UA-Speech vocabulary. The Feature Extraction and Classification pipeline (bottom, green background) extracts representations corresponding to grapheme targets and phonemes via Wav2Vec 2.0 and employs a Siamese MLP to distinguish between *Low* and *High* intelligibility levels.

## 2 Materials and Methods

### 2.1 Datasets Description and Partitioning

We evaluated our method on two publicly available dysarthric speech benchmarks: TORGO [24] and UA-Speech [14].

TORGO includes recordings of dysarthric subjects with cerebral palsy and amyotrophic lateral sclerosis, whose speech impairment severity was assessed via the Frenchay Dysarthria Assessment (FDA) [6]. Audio is captured using both head-mounted microphones and multi-channel arrays, then preprocessed with noise reduction algorithms. In this work, we considered a subset of eight dysarthric speakers with adequate recording quality for feature extraction, yielding approximately 15 hours of speech recordings. Based on their FDA scores, these speakers were grouped into two intelligibility classes (*Low* and *High*) following the same binary scheme adopted for UA-Speech as described later in this subsection. Since TORGO does not provide the percentage intelligibility scores available for UA-Speech, the FDA severity ratings were dichotomised at the functional boundary corresponding to the 50% UA-Speech cut-off: speakers rated severe or moderate-to-severe were assigned to the *Low* class (F01, M01, M02, M04, M05), and speakers rated mild or very mild to the *High* class (F03, F04, M03). This mapping determines only the label inherited by the synthetic samples generated from each reference speaker; any residual imprecision therefore acts as label noise during training rather than as a bias on the reported test metrics, which are computed exclusively on real UA-Speech recordings. Within the scope of this study, the TORGO dataset was used exclusively as a source

of pathological reference audio for the data augmentation pipeline, and was not included in classifier training or evaluation.

UA-Speech collects recordings from 15 individuals affected by dysarthria due to cerebral palsy. The dataset features perceptually-derived intelligibility ratings, which were assigned according to perceptual assessments. For each patient, intelligibility scores were derived from transcriptions provided by five naive listeners, and speakers were labeled into four severity groups, reflecting different levels of speech intelligibility: ‘Very Low’ (0–25%), ‘Low’ (26–50%), ‘Medium’ (51–75%), and ‘High’ (76–100%). The dataset is organized into three blocks, yielding a total of 765 isolated words per speaker.

The classification framework was evaluated using three-fold cross-validation (CV) with training, validation, and test partitions. Crucially, a speaker-independent data separation protocol was strictly enforced across all folds; this ensures that samples from the same speaker never overlap between the training, validation, and test sets, thereby preventing data leakage and ensuring the model’s generalizability to unseen individuals. Furthermore, the original labels of the UA-Speech dataset were aggregated into two macro-classes according to intelligibility: the ‘Very Low’ and ‘Low’ categories were combined to form the *Low Intelligibility* class, while the ‘Medium’ and ‘High’ categories were merged to the *High Intelligibility* class.

## 2.2 Proposed Method

The proposed pipeline (Fig. 1) combines Wav2Vec 2.0-based representations with StyleTTS 2 data augmentation to analyze discriminative cues for dysarthric speech.

**Feature Extraction and Classification Pipeline** The proposed pipeline operates on raw speech signals under a speaker-independent evaluation protocol and leverages Wav2Vec 2.0 exclusively as a feature extractor. The procedure used to obtain grapheme- or phoneme-aligned representations is not part of the proposed method itself, but derives directly from the internal mechanism of the pretrained Wav2Vec 2.0 model. Following the standard configuration described in [1], the input waveform is segmented into 25 ms frames with 5 ms overlap, yielding one frame every 20 ms. During pretraining and fine-tuning, Wav2Vec 2.0 learns an implicit alignment between these frame-level representations and discrete linguistic targets (graphemes or phonemes) through the Connectionist Temporal Classification (CTC) objective [9].

Two distinct public checkpoints are employed, both kept frozen. Grapheme-aligned representations are obtained from `wav2vec2-base-960h` (12 transformer blocks, 768-dimensional states, CTC head over a 32-symbol character vocabulary, fine-tuned on 960 h of LibriSpeech), whereas phoneme-aligned representations are obtained from `wav2vec2-lv-60-espeak-cv-ft` (24 transformer blocks, 1024-dimensional states, CTC head over an IPA phoneme inventory derived

through eSpeak phonemisation). No external forced aligner and no manual segmentation are used. Audio is converted to mono, resampled to 16 kHz, and utterances shorter than 0.3s are discarded; a single forward pass then provides both the frame-level CTC posteriors and the intermediate representations. For a target unit  $u$ , the arg-max token of the CTC posterior is computed at every frame and only the frames whose equals  $u$  are retained, blank frames and frames assigned to other units being discarded. Each retained frame constitutes an independent sample, so the number of samples available for a given unit equals the number of CTC-assigned frames rather than the number of utterances.

Localisation thus relies on the deep, fully contextualised CTC output, whereas the representation that is retained is read at the input of the transformer stack, as this early representation preserves fine-grained acoustic and phonetic information, while deeper layers progressively shift toward more semantic abstractions. Consequently, representations from this stage retain detailed spectral cues, including timbre-related and pathological characteristics that are particularly relevant for intelligibility assessment [4].

These latent representations are subsequently classified via a Multi-Layer Perceptron (MLP) trained in a Siamese fashion [17,3], enabling the estimation of speech intelligibility based on learned similarity relationships.

We extract embeddings for the most frequently occurring linguistic units in the English language [22,15], which also exhibit the highest frequency in the Wav2Vec 2.0 pretraining corpus and match the distribution of letters in the text prompts of the TORGO and UA-Speech datasets. Specifically, we selected thirteen graphemes ('E', 'T', 'A', 'O', 'N', 'I', 'S', 'R', 'D', 'L', 'U', 'F', 'V') and five phonemes ([, tʃ, θ, ð, ɨ]) as representative linguistic units for our classification task. Lower-frequency units were excluded due to insufficient representation within the pathological datasets, which would compromise statistical reliability.

**Data Augmentation Pipeline** The data augmentation pipeline was built upon StyleTTS-2 [16], a text-to-speech system leveraging diffusion-based modeling and style transfer, designed to synthesize new content while preserving the acoustic signature of a reference speaker.

The audio generation procedure required two inputs: reference audio and target text. For each of the eight TORGO speakers, one randomly selected utterance was denoised and used as pathological style reference, so that the augmentation set spans eight distinct pathological styles covering both sexes, both aetiologies represented in the corpus and its full severity range. Preliminary experiments revealed that the TTS system performed poorly when generating isolated words or short-duration audio samples; we therefore opted to synthesize full sentences instead. To preserve the distribution of target linguistic units while producing coherent utterances, we employed the Gemini Pro LLM [26] to generate two collections of 50 and 200 unique sentences, equivalent respectively to 0.73 and 2.91 hours of additional audio. The LLM was conditioned on the UA-Speech word list and instructed to construct sentences incorporating these words along with necessary grammatical connectives, thereby maximizing coverage of the tar-

get vocabulary. Each prompt is synthesised once per reference speaker, yielding  $8 \times 50 = 400$  and  $8 \times 200 = 1600$  synthetic utterances.

To ensure high generation quality, despite the stochastic nature of StyleTTS-2, 15 candidate audio samples were produced for each combination of reference audio and text prompt. The best candidate was selected through a feature similarity comparison process based on the same assumptions used for subsequent classification. We compared the cosine similarity between the reference and candidate clips, computed on the frame-level embeddings of five high-frequency grapheme anchors ('E', 'T', 'A', 'O', 'I'), retaining only the closest sequence and discarding the rest.

The synthetic sentences inherit the intelligibility label from their respective TORGO reference speakers, through the FDA-based mapping described in Section 2.1. It is worth noting that the objective of this augmentation differs from conventional TTS-based approaches aimed at maximizing perceptual audio fidelity: here, synthesis is designed to enrich the representation of specific grapheme and phoneme targets within the Wav2Vec 2.0 embedding space, with the goal of improving unit-level classification rather than reproducing naturalistic dysarthric speech in a clinically rigorous sense. To provide an initial assessment of the clinical plausibility of the generated samples, a subset of the synthetic recordings was informally reviewed by two speech-language pathologists, who confirmed their perceptual resemblance to dysarthric speech at the expected intelligibility levels. The processed synthetic samples were then fed into the Wav2Vec 2.0 feature extractor to perform the classification task. Audio examples of the generated synthetic speech are available online<sup>3</sup>.

### 3 Experimental Setup and Architectures

Regarding feature extraction, the Wav2Vec 2.0 architecture was employed for both tasks, albeit instantiated via distinct model checkpoints. Specifically, weights from separate training regimes were loaded for the graphemes and phonemes, resulting in embedding representations of dimensions 768 and 1024, respectively.

For the classification task, following our previous work [17], a Siamese MLP [3] was adopted. This model learns a similarity metric between vocal embeddings via two weight-shared branches. Each branch processes the input through a dense block, including a linear layer with 256 output units, batch normalization, ReLU activation, and dropout (rate 0.3) [25].

The training focused on minimizing a contrastive loss, where *positive* pairs consisted of samples with matching intelligibility levels and *negative* pairs consisted of samples with different intelligibility levels. Optimization was performed on 100 epochs (with early stopping) using the AdamW algorithm, with a learning rate of  $1 \times 10^{-4}$  and a weight decay of  $5 \times 10^{-2}$  [19].

To provide validation of the proposed feature representations and establish a baseline comparison, a Support Vector Classifier (SVC) was also evaluated under identical experimental conditions as a computationally efficient alternative.

<sup>3</sup> Audio samples: [https://lillobyte.github.io/styletts2\\_dysarthric\\_samples/](https://lillobyte.github.io/styletts2_dysarthric_samples/)

As for the generative phase, specific inference parameters of the StyleTTS-2 were modulated to define the characteristics of the synthesized speech. The selection of the best parameters was driven by the classification performance achieved on a held-out validation subset drawn exclusively from the UA-Speech validation folds, with no overlap with the test set; this procedure therefore preserves the integrity of the speaker-independent evaluation protocol. Specifically, the parameters *alpha* and *beta* govern the style of the generated speech, with *alpha* determining the timbre’s dependence on the reference audio and *beta* controlling the prosody. In this study, *alpha* was set to 0.0 and *beta* was set to 1.0, implying that the prosodic features were derived entirely from the input text model, while an *embedding scale* of 6.0 was applied to the style vectors during the diffusion process to improve the target vocal characteristics.

Consistently with this protocol, synthetic samples are added exclusively to the training partition of every fold, and never to the validation or test partitions, which always consist of real UA-Speech recordings. Augmentation-related leakage is further excluded by construction: the synthetic material is conditioned on TORGO reference speakers, whereas all validation and test data come from UA-Speech, so that no individual can appear both as an augmentation source and in an evaluation partition. The augmentation set is moreover fold-independent, the same material being added to the training partition of each fold, and can therefore not transport information specific to a test fold.

To provide a comparative analysis, two Residual Network architectures, specifically ResNet-50 and ResNet-152 [10], were evaluated as baseline models for the classification task. These models were implemented using the best configuration and framework proposed in our previous research [18], while strictly adhering to the same data partitioning protocols employed for the other experiments described herein. Unlike the proposed approach, ResNet architectures operate on 2D spectrogram representations of the entire utterance, encoding the full acoustic context as a single feature map without isolating individual linguistic units. Their inclusion in this evaluation serves a specific purpose: by comparing utterance-level classification against unit-level classification, it is possible to assess whether targeting specific phonetic or graphemic biomarkers provides a measurable advantage over processing the entire signal. Consequently, the synthetic data generated by the augmentation pipeline was not used to train the ResNet baselines: as described in Section 2, the augmentation strategy is designed to enrich the Wav2Vec 2.0 embedding space at the grapheme and phoneme level, and its benefit is therefore meaningful only for classifiers operating on those representations.

Classification performance was evaluated using the *F1*-score, which is particularly suited for imbalanced datasets as it accounts for both Precision and Recall [5]. To quantify the yield of synthetic samples relative to the added volume, an Augmentation Efficiency Index (AEI) was defined as the ratio between performance gain and percentage increase in data availability per Wav2Vec 2.0 extracted representations. This metric normalizes the evaluation, enabling consistent comparison of augmentation effectiveness despite the non-uniform distri-

**Table 1.** Classification performance ( $F1$ -score: Mean  $\pm$  Standard Deviation) for individual linguistic units. The upper section presents results obtained with the Siamese MLP classifier, under three data configurations. The lower section provides baseline comparisons using ResNet-50, ResNet-152 and SVC

| Linguistic unit        | Original                | Original + Synthetic    |                         |
|------------------------|-------------------------|-------------------------|-------------------------|
|                        |                         | 50 sentences            | 200 sentences           |
| E                      | 71.7% $\pm$ 4.6%        | 76.3% $\pm$ 3.4%        | 75.3% $\pm$ 3.3%        |
| T                      | 64.0% $\pm$ 1.6%        | 70.3% $\pm$ 1.2%        | 67.7% $\pm$ 3.4%        |
| A                      | 69.0% $\pm$ 2.2%        | 73.0% $\pm$ 4.3%        | 74.7% $\pm$ 4.9%        |
| O                      | 63.0% $\pm$ 5.7%        | 69.0% $\pm$ 4.5%        | 70.7% $\pm$ 4.2%        |
| N                      | 64.7% $\pm$ 5.3%        | 67.3% $\pm$ 4.2%        | 69.0% $\pm$ 5.0%        |
| I                      | 62.3% $\pm$ 3.4%        | 71.3% $\pm$ 0.9%        | 71.0% $\pm$ 1.6%        |
| S                      | 65.0% $\pm$ 2.4%        | 69.3% $\pm$ 3.3%        | 68.7% $\pm$ 4.2%        |
| R                      | 65.3% $\pm$ 2.6%        | 76.0% $\pm$ 2.4%        | 75.0% $\pm$ 3.7%        |
| D                      | 62.3% $\pm$ 6.6%        | 70.0% $\pm$ 5.1%        | 71.0% $\pm$ 2.9%        |
| L                      | 58.3% $\pm$ 5.3%        | 63.3% $\pm$ 6.2%        | 65.3% $\pm$ 6.8%        |
| U                      | 63.0% $\pm$ 1.4%        | 72.3% $\pm$ 2.4%        | 73.7% $\pm$ 3.3%        |
| F                      | 59.3% $\pm$ 6.6%        | 64.0% $\pm$ 3.7%        | 68.0% $\pm$ 7.0%        |
| V                      | 63.0% $\pm$ 6.4%        | 66.7% $\pm$ 4.0%        | 74.0% $\pm$ 9.9%        |
| ʃ                      | 58.3% $\pm$ 16.0%       | 68.3% $\pm$ 6.2%        | 68.0% $\pm$ 5.4%        |
| tʃ                     | <b>79.3%</b> $\pm$ 6.2% | <b>80.0%</b> $\pm$ 4.6% | <b>82.7%</b> $\pm$ 8.3% |
| θ                      | 52.7% $\pm$ 2.4%        | 68.3% $\pm$ 2.6%        | 74.7% $\pm$ 6.3%        |
| ð                      | 64.0% $\pm$ 17.7%       | 62.3% $\pm$ 11.1%       | 67.3% $\pm$ 17.3%       |
| ŋ                      | 40.0% $\pm$ 3.6%        | 51.0% $\pm$ 3.7%        | 60.7% $\pm$ 8.2%        |
| average                | 62.5% $\pm$ 5.6%        | 68.8% $\pm$ 4.1%        | 71.0% $\pm$ 5.8%        |
| <b>Baseline Models</b> |                         |                         |                         |
| SVC (avg)              | 63.2% $\pm$ 5.8%        | 68.0% $\pm$ 6.1%        | 70.2% $\pm$ 6.0%        |
| ResNet-50              | 67.7% $\pm$ 5.8%        | -                       | -                       |
| ResNet-152             | 64.4% $\pm$ 7.9%        | -                       | -                       |

**Table 2.** Paired comparisons across augmentation settings. Each observation is one of the 18 linguistic units, scored as its  $F1$  averaged over the three cross-validation folds.  $\Delta F1$  is the mean difference in percentage points, CI the 95% confidence interval of the mean difference,  $p_W$  the exact two-sided Wilcoxon signed-rank test,  $p_t$  the paired  $t$ -test, and  $d_z$  Cohen’s effect size for paired designs

| Comparison                     | $\Delta F1$ | 95% CI        | $p_W$               | $p_t$               | $d_z$ | Impr. |
|--------------------------------|-------------|---------------|---------------------|---------------------|-------|-------|
| Original $\rightarrow$ O+S 50  | +6.31       | [4.24, 8.37]  | $2.3 \cdot 10^{-5}$ | $6.0 \cdot 10^{-6}$ | 1.52  | 17/18 |
| Original $\rightarrow$ O+S 200 | +8.46       | [5.77, 11.15] | $7.6 \cdot 10^{-6}$ | $4.2 \cdot 10^{-6}$ | 1.57  | 18/18 |
| O+S 50 $\rightarrow$ O+S 200   | +2.16       | [0.55, 3.76]  | $1.0 \cdot 10^{-2}$ | $1.2 \cdot 10^{-2}$ | 0.67  | 12/18 |

bution of the extracted Wav2Vec 2.0 representations aligned with grapheme targets and phoneme labels within the datasets. Finally, because frames extracted from the same speaker are not statistically independent, significance testing is performed at the level of the linguistic unit: the fold-averaged  $F1$  of each of the 18

units constitutes one paired observation, which respects the speaker-independent design and avoids the anti-conservative bias that frame-level testing would introduce. On these paired observations we report the Wilcoxon signed-rank test, the paired  $t$ -test, the 95% confidence interval of the mean difference and Cohen’s  $d_z$  effect size.

## 4 Results and Discussion

The primary objectives of this study were to evaluate the impact of the data augmentation pipeline and to identify which grapheme and phoneme targets serve as the most effective acoustic biomarkers for distinguishing between Low and High intelligibility levels.

Table 1 reports the mean  $F1$ -score and standard deviation across the three CV folds for each linguistic unit, as well as the overall average. Results are presented under three configurations: the original UA-Speech data alone, and the original data augmented with 50 and 200 synthetic sentences, respectively. Fig. 2 provides a visual summary of the  $F1$ -score trends across all linguistic units under the three augmentation configurations.

*Statistical Significance.* Table 2 reports the paired comparisons across the three configurations, computed over the 18 units as described in Section 3. Augmentation with 50 synthetic sentences improves  $F1$  by 6.31 percentage points on average, and augmentation with 200 sentences by 8.46 points; both improvements are statistically significant according to both the exact Wilcoxon signed-rank test and the paired  $t$ -test (all  $p < 10^{-4}$ ), with confidence intervals well separated from zero and large effect sizes ( $d_z > 1.5$ ). The difference between the two augmentation volumes is markedly smaller (+2.16 points,  $d_z = 0.67$ ), with only 12 of 18 units improving. Per-unit confidence intervals estimated from three folds are necessarily wide, particularly for the low-frequency phonemes for which few CTC-assigned frames are available; the unit-level observations discussed below should therefore be read as descriptive trends, whereas the statistical claims rest on the aggregate comparisons of Table 2.

Two conclusions follow. The benefit of augmentation is systematic rather than driven by a few favourable units, being observed for 17 and for all 18 units under the two settings; and most of it is already obtained with the smaller augmentation set, the additional 150 sentences yielding a further gain that is both small on average and concentrated on a subset of units. Which units benefit, and why, is the object of the analysis that follows.

*Biomarker Analysis.* Across the full set of units, the affricative phoneme ‘tʃ’ consistently achieves the highest discriminative performance, reaching a baseline  $F1$ -score of 79.3% and 82.7% under full augmentation. This result is consistent with the phonetic properties of the affricate: its production requires a precisely coordinated sequence of complete oral closure followed by a controlled frication release, placing high demands on the neuromotor system [13]. Dysarthric impairment, which disrupts fine motor control, degrades this gesture in a systematic

and measurable way, making it a robust and naturally reliable biomarker. Its high baseline score reflects this intrinsic discriminability, which leaves less room for augmentation-driven improvement.

Cardinal vowels, in particular ‘E’, ‘A’, and ‘U’, also demonstrate consistent discriminative capability, reaching  $F1$ -scores exceeding 73% after augmentation. This is in line with established findings that vowel space reduction and centralization are primary acoustic correlates of dysarthric intelligibility loss [2,28].

Several units that showed limited discriminative power under the original data configuration underwent a substantial improvement with augmentation. The phoneme ‘θ’, with a baseline of 52.7%, reached 74.7% under the 200-sentence configuration, becoming one of the strongest proxies in the set. The grapheme ‘V’ improved from 63.0% to 74.0%, and the phoneme ‘ŋ’ from 40.0% to 60.7%. This pattern indicates that the original dataset contained an insufficient number of tokens for these units, preventing the classifier from learning a stable decision boundary; the synthetic data effectively compensated for this distributional gap. The AEI analysis (Fig. 3) supports this interpretation: ‘θ’ and ‘V’ display the highest augmentation efficiency, while units with strong baseline performance, such as ‘tʃ’, show proportionally lower AEI values, as the margin for improvement is inherently constrained. Conversely, the phoneme ‘ð’ proved the least responsive to augmentation, showing limited improvement (64.0% to 67.3%) coupled with persistently high standard deviation across folds, which may reflect the phoneme’s inherent acoustic variability across dysarthric speakers.

*Effect of Augmentation Volume and Variance.* While the 200-sentence configuration achieves the best overall average  $F1$ -score (71.0%), the improvement over the 50-sentence configuration is not uniform across all units. For several graphemes, including ‘E’, ‘T’, ‘R’, and ‘S’, performance stabilizes or marginally decreases when moving from 50 to 200 synthetic sentences despite the larger augmentation volume. Two concurrent mechanisms are likely responsible for this behavior. First, generating a large number of sentences may introduce modal collapse in the diffusion process of StyleTTS-2, reducing prosodic and timbral variability across samples and causing a subset of synthetic utterances to share similar acoustic artifacts [16]. Second, the Siamese MLP may overfit to these systematic artifacts rather than to pathological features genuinely associated with the dysarthric reference style, reducing generalization on real test data. This composite effect is consistent with the observation that augmentation efficiency is generally higher at 50 sentences for well-represented units, and higher at 200 sentences for units that were severely under-represented in the original data.

This concentration of the gains on the units that the original data represented most poorly, together with the saturation observed between the two augmentation volumes, also clarifies in which sense the benefit is specific to generative augmentation rather than to the mere increase in training-set size. Signal-level augmentation, such as speed or pitch perturbation and spectral masking, transforms the tokens that are already present in the training material, but cannot introduce new tokens of a unit that is rare or absent in the original data, which is

precisely the deficiency the proposed pipeline addresses by synthesising sentences that cover the target vocabulary.

A moderate increase in standard deviation is also observed in the augmented configurations for a subset of units. This reflects the expected behavior in a low-resource, speaker-independent setting: adding synthetic samples from a limited number of reference speakers introduces fold-level variability in the distribution of augmented features, which propagates into higher variance across CV iterations.

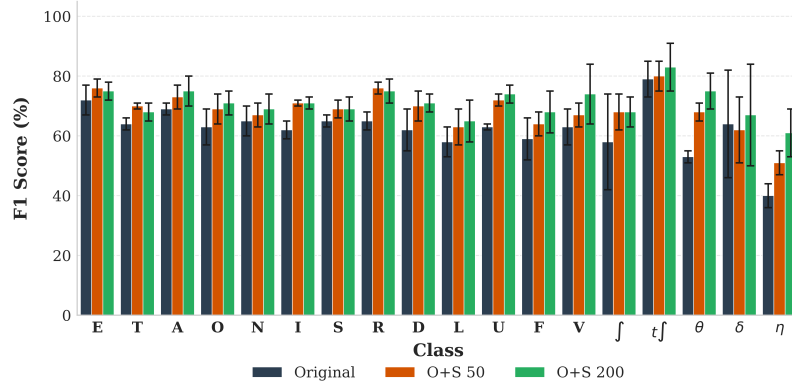


**Fig. 2.** Grouped bar chart comparing  $F1$ -scores across all grapheme and phoneme units under the three augmentation configurations: original UA-Speech data (Original), augmentation with 50 synthetic sentences (O+S 50), and augmentation with 200 synthetic sentences (O+S 200). Error bars represent the standard deviation across the three cross-validation folds.

*Comparison with Utterance-Level Baselines.* The ResNet baselines provide a reference for utterance-level classification: ResNet-50 achieved 67.7% and ResNet-152 achieved 64.4%, as reported in the lower section of Table 1. The lower performance of the deeper architecture is consistent with overfitting on the limited available data [10]. These results support the central hypothesis of this study: classifying at the level of individual graphemes and phonemes yields higher discriminative performance than processing the entire utterance, since the former isolates acoustic regions that are specifically and systematically degraded by dysarthric impairment. Already under the original data configuration, the Siamese MLP operating on the most discriminative units (‘tf’, ‘E’, and ‘A’) individually reaches or exceeds the ResNet-50 score, and units such as ‘R’ and ‘U’ join this group once augmentation is applied: the number of units individually surpassing the ResNet-50 score rises from 3 of 18 to 12 of 18 and 14 of 18 under the two augmented configurations; under the 200-sentence augmentation, the average  $F1$ -score across all units reaches 71.0%, surpassing both baselines.

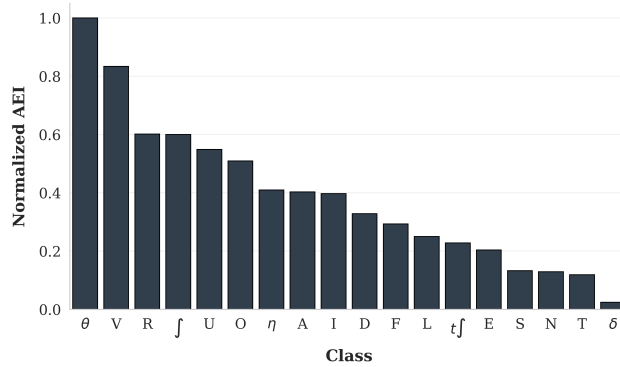


**Fig. 3.** Bar plot displaying the normalized Augmentation Efficiency Index (AEI) for the analyzed graphemes and phonemes. Values are computed on the 200-sentences augmented set.

The SVC evaluated under identical conditions yields consistent trends (63.2%, 68.0%, 70.2%), further corroborating the reliability of the proposed representations across classifiers.

## 5 Conclusion

This study addressed the challenge of data scarcity in automated dysarthric speech intelligibility assessment by proposing a framework that combines SSL with generative data augmentation. Wav2Vec 2.0 was employed to extract contextualized representations aligned with grapheme and phoneme targets, while StyleTTS-2 was used to generate synthetic pathological speech samples conditioned on TORGO reference audio.

Experimental results on the UA-Speech dataset show that unit-level classification consistently outperforms utterance-level approaches, with the affricative ‘tf’ and cardinal vowels emerging as the most discriminative biomarkers across all configurations. Synthetic data augmentation improved classification performance for the majority of linguistic units, with the largest gains observed for units that were under-represented in the original dataset. At the same time, the results indicate that augmentation volume should be calibrated per unit: for well-represented units, the 50-sentence configuration is sufficient, while severely under-represented units benefit from the full 200-sentence set.

A limitation of the current framework concerns the labeling of synthetic samples, whose intelligibility labels are inherited from the TORGO reference speakers without formal perceptual validation. The informal review conducted by two speech-language pathologists provides initial support for the clinical plausibility of the generated speech; however, a structured perceptual evaluation by clinical experts remains an open methodological step.

Future work will extend this framework to multi-class severity assessment and explore its applicability to other motor speech disorders. Furthermore, a formal perceptual validation study of the generated samples, in the form of a Mean Opinion Score evaluation conducted by a panel of speech-language pathologists, remains as future work. A systematic comparison against alternative augmentation strategies (signal-level perturbation, a neutral-style synthesis control and other generative architectures) evaluated at matched added-data volume through an identical extraction and classification pipeline, is likewise required to quantify how much of the benefit is specific to StyleTTS-2, and constitutes the natural continuation of the present work. Additional investigation will address the evaluation of other SSL models, the combination of grapheme- and phoneme-aligned representations to exploit complementary acoustic information, and the design of phonetically balanced text prompts in collaboration with speech-language pathologists to maximize the coverage and discriminative value of the generated data.

**Acknowledgments.** This work was supported by the project “Talia: intelligent vocal biomarkers for pathology early detection and follow-up Assessment” (Ministerial Decree 31/12/2021, Agreements for Innovation, Horizon Europe – Pillar II; Decree no. 0352746 of 13 October 2023; Project No. F/350215/01-04/X60).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Code and Data Availability

The code implementing the feature extraction procedure, the data augmentation pipeline and the classification experiments reported in this work is publicly available at:

<https://github.com/LilloByte/AII-2026-code-and-data-Dysarthric-phonemes-classification->

The synthetic dysarthric speech generated for the augmentation experiments, comprising both the 50- and the 200-sentence sets together with a manifest reporting, for each file, its TORGO reference speaker, inherited intelligibility label and duration, is available at:

[https://univpm-my.sharepoint.com/:f:/g/personal/s1117688\\_pm\\_univpm\\_it/IgA1rFOL5urCTqTa813vgBbVAYz57tAc6UawAdtIilo2ix0?e=JbqnMl](https://univpm-my.sharepoint.com/:f:/g/personal/s1117688_pm_univpm_it/IgA1rFOL5urCTqTa813vgBbVAYz57tAc6UawAdtIilo2ix0?e=JbqnMl)

The TORGO and UA-Speech corpora are not redistributed here: access to them must be requested from their respective owners.

## References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020)
2. Bang, Y.I., Min, K., Sohn, Y.H., Cho, S.R.: Acoustic characteristics of vowel sounds in patients with Parkinson disease. *NeuroRehabilitation* **32**(3), 649–654 (2013)
3. Chen, X., He, K.: Exploring simple siamese representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 15750–15758 (2021)
4. Chernyak, B.R., Bradlow, A.R., Keshet, J., Goldrick, M.: A perceptual similarity space for speech based on self-supervised speech representations. *The Journal of the Acoustical Society of America* **155**(6), 3915–3929 (2024)
5. Diallo, R., Edalo, C., Awe, O.O.: Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, mcc, and f1 score. In: *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*, pp. 283–312. Springer (2024)
6. Enderby, P.: Frenchay dysarthria assessment. *International Journal of Language & Communication Disorders* **15**(3), 165–173 (1980). <https://doi.org/https://doi.org/10.3109/13682828009112541>, <https://onlinelibrary.wiley.com/doi/abs/10.3109/13682828009112541>
7. Fahed, V.S., Doheny, E.P., Collazo, C., Krzysztofik, J., Mann, E., Morgan-Jones, P., Mills, L., Drew, C., Rosser, A.E., Cousins, R., et al.: Language-independent acoustic biomarkers for quantifying speech impairment in Huntington’s disease. *American Journal of Speech-Language Pathology* **33**(3), 1390–1405 (2024)
8. Favaro, A., Moro-Velázquez, L., Butala, A., Motley, C., Cao, T., Stevens, R.D., Villalba, J., Dehak, N.: Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson’s disease. *Frontiers in Neurology* **14**, 1142642 (2023)
9. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd international conference on Machine learning*. pp. 369–376 (2006)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
11. Huang, A., Hall, K., Watson, C., Shahamiri, S.R.: A review of automated intelligibility assessment for dysarthric speakers. In: *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. pp. 19–24. IEEE (2021)
12. Joshy, A.A., Rajan, R.: Automated dysarthria severity classification using deep learning frameworks. In: *2020 28th European signal processing conference (EU-SIPCO)*. pp. 116–120. IEEE (2021)
13. Kent, R.D., Kim, Y.J.: Toward an acoustic typology of motor speech disorders. *Clinical linguistics & phonetics* **17**(6), 427–445 (2003)
14. Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J.R., Huang, T.S., Watkin, K.L., Frame, S.: Dysarthric speech database for universal access research. In: *Interspeech*. vol. 2008, pp. 1741–1744 (2008)
15. Lewand, R.: *Cryptological Mathematics*. Mathematical Association of America (11 2000). <https://doi.org/10.1090/clrm/016>

16. Li, Y.A., Han, C., Raghavan, V., Mischler, G., Mesgarani, N.: StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems* **36**, 19594–19621 (2023)
17. Lillini, D., Aironi, C., Migliorelli, L., Gabrielli, L., Squartini, S.: SiCRNN: A siamese approach for sleep apnea identification via tracheal microphone signals. *Sensors* **24**(23), 7782 (2024)
18. Lillini, D., Aironi, C., Migliorelli, L., Gabrielli, L., Squartini, S.: Dysarthric speech classification: A comparative analysis of decision-support methods. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (2025). <https://doi.org/10.1109/IJCNN64981.2025.11228715>
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Maier, A., Schuster, M., Batliner, A., Nöth, E., Nkenke, E.: Automatic scoring of the intelligibility in patients with cancer of the oral cavity (2007)
21. Millet, J., Dunbar, E.: Do self-supervised speech models develop human-like perception biases? arXiv preprint arXiv:2205.15819 (2022)
22. Mines, M.A., Hanson, B.F., Shoup, J.E.: Frequency of occurrence of phonemes in conversational english. *Language and Speech* **21**, 221 – 241 (1978), <https://api.semanticscholar.org/CorpusID:43625690>
23. Pasad, A., Chou, J.C., Livescu, K.: Layer-wise analysis of a self-supervised speech representation model. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 914–921. IEEE (2021)
24. Rudzicz, F., Namasivayam, A.K., Wolff, T.: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation* **46**(4), 523–541 (2012)
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**(1), 1929–1958 (2014)
26. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
27. Tripathi, A., Bhosale, S., Kopparapu, S.K.: A novel approach for intelligibility assessment in dysarthric subjects. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6779–6783. IEEE (2020)
28. Turner, G.S., Tjaden, K., Weismer, G.: The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research* **38**(5), 1001–1013 (1995)
29. Venugopalan, S., Shor, J., Plakal, M., Tobin, J., Tomanek, K., Green, J.R., Brenner, M.P.: Comparing supervised models and learned speech representations for classifying intelligibility of disordered speech on selected phrases. arXiv preprint arXiv:2107.03985 (2021)
30. Yeo, E.J., Choi, K., Kim, S., Chung, M.: Cross-lingual dysarthria severity classification for English, Korean, and Tamil. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 566–574. IEEE (2022)