Article:
Leung, W.-Z., Christensen, H. and Goetze, S. (2026) Towards automating the Frenchay dysarthria assessment: Can neural phoneme posteriorgrams inform the analysis of dysarthric speech? Speech Communication, 179. 103379. ISSN: 0167-6393

https://doi.org/10.1016/j.specom.2026.103379

Towards Automating the Frenchay Dysarthria Assessment: can neural phoneme posteriorgrams inform the analysis of dysarthric speech?

Wing-Zin Leung[a], Heidi Christensen[a], Stefan Goetze[a,b]

[a]*Speech and Hearing (SpandH), School of Computer Science, The University of Sheffield, UK*
[b]*South Westphalia University of Applied Sciences, Iserlohn, Germany*

Abstract

Dysarthria is a type of motor speech disorder that reflects abnormalities in motor movements required for speech production. In clinical practice, identifying characteristic signs and symptoms of the neuropathophysiology underlying a dysarthria is vital for diagnosis and management. The gold standard for dysarthria assessment is auditory-perceptual evaluation by a speech and language therapist for differential diagnosis and management decisions. As the process is time-consuming for clinicians, there is growing interest in automatic dysarthria assessment (ADA). Recent approaches to ADA primarily focus on the classification of broad intelligibility or speech severity labels. However, this does not have much clinical utility and the assessment of communication-relevant parameters do not distinguish between dysarthria types and pathomechanisms. Studies on the classification of dysarthria function or clinical test protocol scores focusing on aspects of dysarthric speech production (such as the Frenchay dysarthria assessment (FDA)) are limited.

Therefore, this paper focuses on the preliminary steps towards clinically interpretable ADA, including automatic FDA assessment. The phoneme posteriorgram (PPG) is a time-varying categorical distribution over acoustic speech units, and recent work demonstrates interpretable speech pronunciation distance for downstream tasks, e.g. pronunciation reconstruction. This work extends recent advances in posterior-based phoneme research and mispronunciation models to dysarthria assessment, exploring the extent to which dysarthric speech features in the FDA (identified by auditory-perceptual evaluation in clinical practice) are captured by PPG information. To achieve this, FDA aspects are systematically evaluated. The results show that interpretable PPG probability can capture dysarthric speech features that are related to motor system dysfunction.

Keywords: Automatic dysarthria assessment, dysarthric speech features, phoneme posteriorgram

1. Introduction

Dysarthria is a type of motor speech disorder (MSD) resulting from disturbances in muscular control over speech production due to lesions of the central or peripheral nervous system [1]. As the patient group is heterogeneous, identifying characteristic signs and symptoms of the underlying neuropathophysiology is vital for diagnosis and management [2]. In clinical practice, auditory-perceptual evaluation of dysarthria by a speech and language therapist (SLT) is the gold standard for differential diagnosis, severity judgement, and management decisions [3, 4]. An SLT's assessment methods can vary from a clinical test protocol focusing on aspects of speech production (e.g. to diagnose dysarthria type, and aid in neurological diagnosis in some cases) to conversational interaction (e.g. for detecting the presence of a neurogenic speech disorder) [5]. In the former case, the Frenchay dysarthria assessment (FDA) [6] is widely used in clinical practice as a criterion-referenced and replicable clinical test for dysarthria diagnosis [7], with sensitivity to changes in the pattern of speech behaviour [6]. However, auditory-perceptual analysis is often time-consuming as it relies on comprehensive evaluation by healthcare professionals; therefore, there is increasing interest in and research into automatic (computer-based) dysarthria assessment.

Recent research on automatic dysarthria assessment (ADA) focuses on dysarthric speech intelligibility or speech severity classification (as well as dysarthria vs. control speaker classification [8, 9] and disorder classification [10]). The commonly used *TORGO* dataset [11] includes FDA assessment scores and studies primarily classify labels based on FDA sentence intelligibility ratings to either classify speech intelligibility [12] or use intelligibility ratings as an indicator of speech severity[1] [23, 24, 25, 26, 27].[2] However, *TORGO* classification studies often lack detail in specifying the classification labels

[1]Speech intelligibility is a common clinical metric for communication effectiveness in MSDs [13, 14], and is often used as an indicator of International Classification of Functioning, Disability and Health (ICF) [15] Activity and a proxy for speech severity [16, 17]. However, clinical guidelines [18, 19] identify the need for a holistic approach (e.g. the primary goal of intervention should also maximise efficiency and naturalness of communication in context of the ICF [1, 14]), and recent research is increasingly advocating for broader holistic approaches to speech severity, including impact, perception and communicative participation [20, 21]. Therefore, although communication-relevant parameters are fundamental in holistic care and provide information on functional outcome goals and measures [22], current approaches to ADA (i.e. the classification of a single broad label per speaker in isolation) have limited clinical utility beyond broad estimators of speech intelligibility.

[2]The majority of these studies use *TORGO* in combination with the UASpeech (UAS) dataset and correspond intelligibility ratings between these datasets using percentage thresholds. However, there are discrepancies includ-

used or the derivation of classification groups [29, 30], or use groups that do not correspond to any FDA scores [31, 32]. [33] classifies a total aggregate FDA score for the Mandarin Subacute Stroke Dysarthria Multimodal dataset. However, this presumes that the FDA sections are weighted equally (some FDA sections have much greater impact on intelligibility and communicative effectiveness/efficiency [34]) and the method loses clinical diagnostic and management detail. The Nemours database [35] contains both FDA scores and scores based on the percentage of words correctly identified in nonsense sentences by naive listeners, and studies primarily classify labels based on the percentage of correct words identified as an indicator of speech severity [36, 37]. Classification studies on other dysarthric datasets use and classify intelligibility labels, for example based on orthographic transcription, such as the prediction of intelligibility labels determined by naive listener transcription scores from the commonly used UAS database [38, 39] or SLT transcription from the quality of life technology (QoLT) database [40, 41]. However, these approaches lack details that are clinically salient such as (sub-)scores that provide information relating to speech production and the formation of a clinical dysarthria profile. In general, the assessment of communication-relevant parameters (e.g. speech intelligibility and speech severity) do not distinguish between dysarthria types and pathomechanisms (and the involved motor systems). This paper addresses this by investigating the classification of dysarthric speech features from the FDA that are used to categorically diagnose dysarthria and to compile results that are applicable to the clinical management of dysarthria [6].

Current state of the art (SoTA) approaches to ADA broadly use two approaches: (i) handcrafted features with a focus on interpretability [42], generally in combination with classical machine-learning classifiers (e.g. support vector machines (SVMs) [31] or Random Forest classifiers [43] with Mel-frequency cepstral coefficient (MFCC) [44, 23], glottal-based [45], sparsity-based [46], openSMILE [9] or eGeMAPS [25] features) and (ii) deep learning approaches to automatically extract discriminative features, commonly with e.g. convolutional neural networks (CNNs) [47, 48], long short-term memory (LSTM) [24], or autoencoders [10]. There is also growing research using self-supervised representation (SSR) features (e.g. wav2vec 2.0 (W2V2) [49] or HuBERT [25]).

Increasingly, however, research highlights the importance of interpretable features and models and there is growing concern that classifiers are learning other aspects of audio samples instead of dysarthric speech features. [8] shows that the majority of SoTA approaches achieve the same or even significantly better performance using only non-speech segments, indicating that classification approaches validated on commonly used datasets like the *TORGO* [11] and *UAS* [38] databases are rather learning characteristics of the recording environment. [50] utilise speaker identity-invariant features (containing discriminative information for Parkinson's disease (PD) classification to address concerns that models are learning speaker identity information. Also, studies have focused on interpretability, e.g. interpretable model layers [51], model constraints and interpretable acoustic [52] and phonological features [53].

Previous work on more clinically relevant automatic systems, like the classification of dysarthria function or the prediction of FDA scores, is limited. [54] uses spectral and respiratory pressure features to predict lip and laryngeal scores to characterise and differentiate ataxic and mixed dysarthric speech features, and [33] uses vowel space diagrams (based on audio-visual information) to predict FDA scores, and demonstrate feature interpretability. The relationship between quadrilateral vowel space area (VSA) and perceptual dysarthric intelligibility has received relatively more focus. Generally, studies have concluded that dysarthric speakers have reduced VSA relative to control speakers (indicative of decreased articulatory working space and less perceptual vowel distinction [55]). Recent studies have addressed limitations of interpreting dysarthric speech intelligibility from VSA (e.g. the articulatory working space being inferred from corner vowels [56], or high dialect/accent impact [57]) by implementing alternative measures (based on vowel space) to quantify vowel distinction [56, 58].

Finally, phoneme recognition systems [59, 60], mispronunciation detection of speech production errors [61] as well as posterior-based phoneme confidence scores [62] have made significant progress in recent years. Goodness of Pronunciation (GoP) models have been well established in non-native speech pronunciation assessment [63], and recent studies have verified their use in speech disorder assessment [64] (e.g. using MFCC [64] and SSR [65] features). A representation of the probability of mispronunciation is calculated with the GoP algorithm [66], and outlier detection can be performed to quantify the level of deviation from control speech [65]. However, these approaches remain limited in identifying dysarthric speech features and creating an overall dysarthria profile (and therefore remain limited in clinical application).

This paper proposes the use of PPG-based analysis to address these issues. A PPG is a time-varying categorical distribution over acoustic speech units, e.g. phonemes [67] (cf. Section 2.2 for details), and PPGs will be used as an interpretable feature in this work to evaluate dysarthric speech features and FDA scores. PPGs have been used for dysarthric voice conversion [68, 69], a pronunciation error metric for generated dysarthric speech [70], and as a classification model feature (e.g. SVM models and binary classification of control and dysarthric speakers [71]). Recent work has investigated the utility of PPGs for downstream tasks, including interpretable speech pronunciation distance [72], however, this has not previously been explored with MSDs. This paper extends recent advances in posterior-based phoneme research and mispronunciation models by using interpretable PPG speech pronunciation distance to evaluate dysarthric speech features that are determined by auditory perceptual evaluation during clinical dysarthria assessment, including FDA scoring. The contributions of this paper can be summarised as follows:

- This paper introduces a framework for the evaluation of interpretable features (based on analysis of dysarthric speech

ing whether listeners are clinical or non-clinical, and the methods to quantify intelligibility and the threshold ranges assigned to each category [21, 22, 28].

features across relevant FDA aspects) to address the lack of research on the automation of FDA score prediction.

- Previous approaches to dysarthria classification have primarily focused on broad scores based on the deviation between control and dysarthric speech, or the classification of intelligibility labels (or intelligibility as an indicator of speech severity). This falls short of what is done in current clinical gold-standard dysarthria assessment, where auditory-perceptual evaluation is routinely conducted to define speech features related to dysfunction of the motor system (and the prosodic consequences).

- FDA aspects are systematically evaluated (including manual listening by a SLT as appropriate). A detailed analysis of the auditory perceptual features relevant to FDA scoring in the *TORGO* have not been previously documented, as well as analysis of the dysarthric speech processes in context of an interpretable feature. As a preliminary step towards clinically interpretable ADA (including automatic FDA assessment), this paper shows that interpretable PPG probability can capture these dysarthric speech features, with the potential utility to classify speech production impairment and a dysarthria profile.

The remainder of the paper is structured as follows: Section 2 outlines the methodology, including introduction of the *TORGO* dysarthric dataset which will be used for the experimental work, the neural PPG model, and the features used for analysis. The evaluation of PPG features and the control speaker data is presented in Section 3, and Section 4 shows the results of the dysarthric data. Finally, Section 5 concludes the paper.

2. Methodology

In gold standard clinical practice, auditory perceptual evaluation is conducted by qualified health professionals to diagnose and rate the severity of dysarthric speech features. For FDA assessment, dysarthric speech features (identified by auditory perceptual evaluation) are used to categorically diagnose the dysarthria and compile results that are applicable to the clinical management of the dysarthria [6]. In this work, FDA aspects are systematically evaluated, and recent advances in posterior-based phoneme research are applied to dysarthric speech assessment by analysing the extent to which PPG features can capture these dysarthric speech features. The *TORGO* database as data source is introduced in Section 2.1. The *TORGO* is commonly used for studies on dysarthric speech, and contains phoneme alignment data and FDA assessment scores. The *UAS* dataset is also commonly used but does not contain FDA scores and therefore has not been used in this study. The PPG model and the PPG features used in the study are briefly introduced in Section 2.2. PPG evaluation of the *TORGO* control and dysarthric data are reported in Sections 3 and 4, respectively.

2.1. *The* TORGO *database*

The *TORGO* database, frequently used in dysarthric speech research, contains 15.18 hours of acoustic data, 5, 184 phoneme alignment files, and 3D articulatory feature data [11]. Participants were recruited from a Research Institute in Canada, and data was gathered from 8 dysarthric speakers with a diagnosis of amyotrophic lateral sclerosis (ALS) (spastic & ataxic dysarthria) or cerebral palsy (CP) (flaccid & spastic dysarthria) and 7 age-gender-matched control speakers. Acoustic data was recorded through 2 microphones: an 8-channel array microphone and a head-mounted close-talk microphone. All speakers had the same prompts, resulting in a large overlap in word and sentence prompts between speakers [73]. Corrupted recordings, utterances with no transcription or non-speech transcription, and utterances that are too short to contain speech (audio with duration < 0.4 seconds) were discarded [74] (leaving $16,514$ processed audio files with 13.54 hours duration).

2.1.1. *Frenchay Dysarthria Assessment ratings for TORGO data*

The dysarthric speakers in the *TORGO* data were assessed by a SLT using the FDA [6]. The FDA is divided into 8 sections relating to aspects of speech function. For each FDA aspect, there are subtasks that either require physical examination/observation (e.g. tongue at rest), examination of oromotor movements (e.g. lip seal, tongue elevation), or perceptual rating during connected speech tasks or conversation (e.g. note nasal resonance and nasal emission during spontaneous conversation). For each subtask there is a 5-point descriptive scale (between 'A' (within normal limits) to 'E' (no function), where intermediate gradings can be selected, e.g. 'A/B' being between 'A' and 'B') that are rated by the clinician to assign a subscore. There is also a section to record factors that may influence speech, for example hearing difficulties, issues with dentures, and posture.

To compile results, subscores are recorded on a bar graph to visualise the severity ratings achieved on subtasks and highlight the pattern of speech disorder for diagnosis and management. Figure 1 shows the visualised FDA scores for speaker M01 where FDA aspects which are unaffected (reflex, jaw) and those that are more severely affected (respiratory, lips, laryngeal, tongue) are distinguished. The FDA requires that clinicians analyse the behaviour of each aspect in isolation to examine relative abilities and disabilities [34]. As speech is a unitary system, impairment in one aspect can impact on another (e.g. respiration control impairment can cause poor laryngeal function and weak plosives [75]), and the FDA test procedure facilitates the localisation of aspects and speech features.

The speech pattern and profile can indicate the level and type of neurological dysfunction due to damage in the central or peripheral nervous system, and the identification of abnormal movement (as well as retained movements) is required to distinguish between dysarthria subtypes (flaccid, spastic, ataxic, hypokinetic, hyperkinetic, and unilateral upper motor neuron (UMN)) [1]. The Darley, Aronson, and Brown (DAB) clinical taxonomy for MSDs [76] distinguishes clusters of atypical

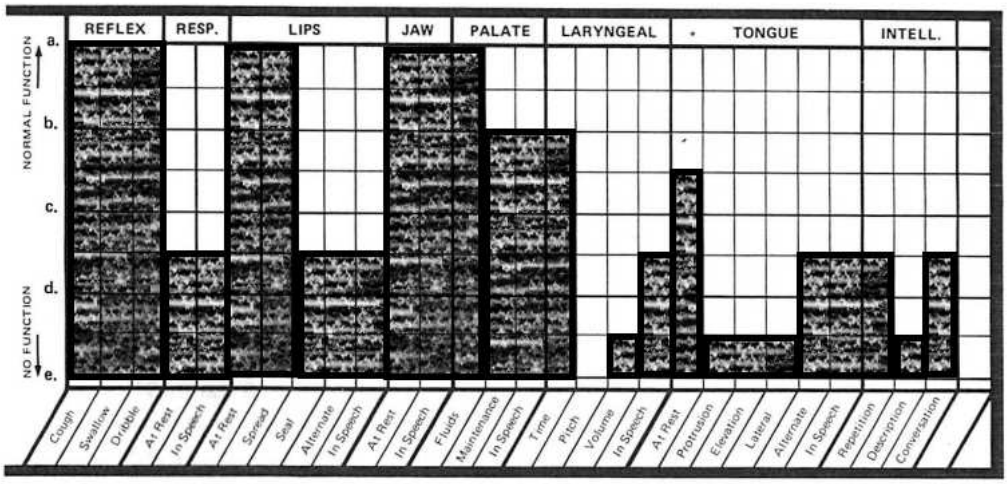

Figure 1: A bar graph visualising FDA subscore ratings for speaker M01. Adapted from [6].

speech features that are associated with both primary and mixed types of dysarthria [77], with studies validating the classification of dysarthria type and characteristics by underlying pathology [78]. For example, flaccid dysarthria can have focal (isolated) areas (and therefore aspects) of involvement depending on lower motor neurons affected (and specific muscle groups innervated by affected cranial nerves) with deviant speech characteristics related to muscle weakness and hypotonia (reduced tone), in contrast to spastic dysarthria with impaired movement patterns (often for all components of the speech subsystem, albeit not equally) with deviant speech characteristics related to excess muscle tone reflective of the central nervous system motor pathways [1].

The FDA does not include instructions for calculating a total score. As highlighted, a profile of dysarthria (i.e. the pattern of speech disorder) is clinically useful for diagnosis and management, and it is the FDA aspects and subscore severities that are standardised and norm-referenced [34]. A total score would assume that all aspects and subtasks have the same weighting, and that dysarthria types will impact across subtasks equally. In summary, it is important to determine the pattern of speech disorder that underlies the broad intelligibility or severity labels that are often classified in ADA studies, and the FDA provides a standardised assessment framework to isolate speech aspects to compile a dysarthria profile.

The lip, jaw, soft palate, laryngeal and tongue aspects include 'in speech' subtasks that are perceptually rated during conversation, and these ratings for the *TORGO* speakers are displayed in Table 1. It is to be noted that the jaw aspect was omitted from the second edition of the FDA (FDA-2) as clinicians rarely observed jaw abnormality and jaw information did not assist diagnosis [34]. The FDA reflex aspect is not included, as the subscores are rated on cough and swallow (during eating and drinking), and dribble (presence of drooling) which cannot be assessed by audio-only methods.[3] The analysis in this paper will focus on the 'in speech' subscores from the FDA.

It can be observed that speakers F01 and M04 in the TORGO data both have an FDA sentence intelligibility subscore rating of 'D/E' (a score between 16.67% and 41.67% on a sentence interpretation task),[4] but there are differences in severity of subscore ratings that characterise the dysarthria (and underlie the dysarthric phonological processes that impact on intelligibility). Furthermore, speakers F03 and F04 both have 'A' ratings for all 'in speech' subscores, but have more severe ratings in e.g. cough and drooling reflex, the observed tongue at rest, and lingual oromotor movements (e.g. protrusion, elevation and lateral movement). These examples highlight that a dysarthria profile is required for clinically meaningful results.

Table 1: FDA 'in speech' subscores rated on a 9-point scale between 'A' (normal function) to 'E' (no function) with 'A/B' being between 'A' and 'B'. 'F' and 'M' denote gender, and the numeral the participant number in the dataset [11]. FDA aspects with an asterisk(*) are analysed in this paper.

	F01	F03	F04	M01	M02	M03	M04	M05
Respiration	C	A	A	C/D	C/D	A/B	C	C/D
Jaw	C/D	A	A	A	A	A	C	A
*Soft Palate (Section 4.1)	D	A	A	B	B	A	A	B
*Tongue (Section 4.3)	D	A	A	C/D	C/D	A	D	C/D
*Lips (Section 4.4)	C	A	A	C/D	C/D	A/B	C	C/D
*Laryngeal (Section 4.5)	D	A	A	C/D	C/D	A/B	E	C/D
*Intel. (sentences) (Section 4.6)	D/E	A	A	D/E	D/E	A	D/E	C

2.1.2. *TORGO Acoustic Phoneme-Aligned Data*

For the analysis conducted in this paper, the *TORGO* phoneme alignment data is used in combination with the processed audio data described in Section 2.1. As outlined below,

[3]There is growing research predicting swallow function and aspiration risk from audio signals e.g. aspiration risk from voice features [38] or swallow performance from audio signals (of e.g. cervical auscultation [79]). However, the reflex aspect is rated on a cough (coughing/choking and degree of difficulty clearing the throat during eating and drinking), swallow (degree of oropharyngeal dysphagia as determined by clinical swallowing evaluation [80] and fluid/diet modification and non-enteral feeding status) and dribble/drool (rated on observation of drooling, including when e.g. drinking, concentrating, at rest) subscores.

[4]Note that the sentence stimuli and score boundaries have been updated in the FDA-2 [34].

significant pre-processing steps are required to ensure accuracy of timestamps, annotation transcripts, and audio and alignment file pairing, and therefore the code will be made publicly available to allow reproduction of this paper's research[5]. Annotations in the *TORGO* phoneme alignment data include some word and syllable repetition (e.g. /wi wi wɝ/ and /bʊ bʊ/, respectively), and some insertions (e.g. /aɪ lʊkt ʌp/ to /haɪ lʊkt ʌp/), substitutions (e.g. /sleɪ/ to /fleɪ/) and deletions, but the annotations are not comprehensive. [11] provides an analysis of the substitution and deletion errors in the *TORGO*, and therefore the phoneme alignment data will be analysed in the context of this. The phoneme alignments in the *TORGO* are generated from one microphone (90.71% of alignment files correspond to the headset microphone), and for some sessions the microphone offset for the other microphone is available but there are inconsistencies that would require manual annotation to validate. Therefore, only the microphone that corresponds to the phoneme alignments is used in this work to ensure accuracy. The phoneme alignment data is processed by correcting phoneme token typos, removing samples with no phoneme alignment and incorrect alignments (e.g. that begin after the length of the sample). After pre-processing, there is phoneme-aligned acoustic data for 9 of the *TORGO* speakers, resulting in 1683 control speaker and 1612 dysarthric speaker audio samples remaining (cf. Table 2 for details).

Table 2: TORGO Phoneme Aligned Data

	F01	F03	F04	M01	M04	M05	MC01	MC02	MC04
Wav files	118	516	228	90	252	408	707	360	616
% total	100	94.68	93.83	24.26	64.78	87.18	97.52	96.77	62.16
Duration (mins)	4.90	21.04	13.35	7.13	19.39	31.45	36.68	19.20	21.71
Single word	98	392	163	70	195	315	533	281	468
Multi word	20	124	65	20	57	93	174	79	148
Phoneme tokens	1114	6179	3216	1275	3565	5514	8800	4137	7328

The 'Wav files' row shows the total number of processed audio samples (that correspond to the microphone in the phoneme alignment data), and the '% total' row shows the percentage of these wav files remaining after phoneme alignment data processing (80.60% and 75.74% of control and dysarthric speaker single microphone audio samples, respectively). Finally, note that phoneme classes are not balanced in *TORGO*, and there is an inadequate volume and variety of data to select a phoneme-balanced subset. The /dʒ/ and /ʊ/ phonemes are under-represented in the dataset (and for some speakers there are no instances of these phonemes in the aligned data), and /θ/ is under-represented in the dysarthric data. Therefore, /dʒ/ and /ʊ/ will be excluded from analysis, and /θ/ will be excluded from the analysis of dysarthric data.

2.2. The High-Fidelity Neural (H-FN) Phoneme Posteriorgram (PPG) model

The High-fidelity Neural (H-FN) PPG model[6] architecture is composed of an input convolution layer, five Transformer layers with self-attention and a feed-forward network, followed by an output convolution layer with a softmax activation function (that produces a categorical distribution over 39 ARPAbet phoneme tokens from the Carnegie Mellon University (CMU) Pronunciation Dictionary[7], and a non-speech token) [72]. The model is pretrained on typical speech (Common Voice (CV) 6.1 dataset [81]) and evaluated on a held-out partition of CV, the CMU Artic [82] and TIMIT [83] datasets). [72] demonstrates interpretable distance (of framewise pronunciation error) on typical speech validation data, showing relatively higher probability between e.g. corresponding voiced and unvoiced fricatives.

The H-FN PPG models were trained on both Mel spectrogram and W2V2 [49] feature representations. For this work, PPG models with Mel spectrogram and W2V2 input features were compared (not shown), and the results indicate that the Mel spectrogram model generalised better to control speech in the *TORGO* (cf. Section 3 for details on the evaluation methodology), and therefore the results of this model are reported. The Mel spectrogram features are log-energy magnitude spectrograms with 80 channels computed from raw audio with a hop size of 10 ms and a window size of 64 ms at a sampling frequency of $f_s = 16$ kHz.

2.2.1. Phoneme Posteriorgram

Inference is performed for each wave file with the H-FN PPG model to output PPGs of dimension $|\mathcal{P}| \times T$, where $|\mathcal{P}|$ denotes the cardinality of the phoneme token set $\mathcal{P}$ (with $|\mathcal{P}| = 40$, according to [72]), providing a categorical distribution over phoneme tokens for every frame in the audio. The posteriorgram can be written as a matrix $\mathbf{G} \in \mathbb{R}^{|\mathcal{P}| \times T}$, where each entry $g_{p,t}$ denotes the inferred probability that the phoneme token p is present in frame t.

2.2.2. Phoneme Aligned Posteriorgram

Each PPG frame in $\mathbf{G} \in \mathbb{R}^{|\mathcal{P}| \times T}$ represents a time step of 10 ms (160 samples per frame shift). Phoneme alignment PPG windows are extracted using the *TORGO* alignment data and averaged over T_{p_i} frames t' for each phoneme occurrence i in the (TORGO) phoneme alignment data to calculate an average distribution which can be expressed by the vector

$$\bar{\mathbf{g}}_{p_i} = \frac{1}{T_{p_i}} \sum_{t'=1}^{T_{p_i}} \mathbf{g}_{p_i} \tag{1}$$

of size $|\mathcal{P}|$ for each aligned phoneme p_i present in T_{p_i} frames.

For each phoneme p, the probability assigned to each other possible phoneme token q in the distribution $\bar{\mathbf{g}}_p$ can be modelled by $P(q|p) = \bar{\mathbf{g}}_{p,q}$, where $P(q|p)$ denotes the probability of predicting phoneme estimate q when the (TORGO) phoneme aligned annotation is phoneme p. Thus, the correct probability (i.e. the probability of the correct phoneme token) is when $q = p$, and the misclassification probability is the case when

[5] The code to process the TORGO phoneme aligned dataset can be found at: https://github.com/WingZLeung/TORGO_aligned.

[6] Neural PPG model: https://github.com/interactiveaudiolab/ppgs.

[7] CMU Pronunciation Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

$q \neq p$ (where $q|p$ refers to the probability of q when the annotation label is p). The phoneme q with the highest probability (i.e. arg $\max_{q \in \mathcal{P}} P(q|p)$) is the model's most likely prediction for the phoneme.

2.2.3. Acoustic Phoneme Similarity

In [72], interpretable acoustic similarity is demonstrated for typical speech between phonemes using the class weighted probability assigned to phoneme q when phoneme p is the ground truth label. For a given speaker, the average distribution over all instances of a given phoneme is calculated by

$$\hat{\mathbf{g}}_p = \frac{1}{N_p} \sum_{i=1}^{N_p} \bar{\mathbf{g}}_{p_i}, \tag{2}$$

where $\hat{\mathbf{g}}_p$ represents the overall averaged distribution for phoneme p across all its instances, N_p denotes the total number of instances of phoneme p, and $\bar{\mathbf{g}}_{p(i)}$ is the mean-pooled distribution for the i-th instance of phoneme p. Acoustic similarity can be computed by comparing the probability allocated to phoneme q when phoneme p is the phoneme annotation label. The overall averaged distribution $\hat{\mathbf{g}}_p$ is calculated for all phonemes in the phoneme token set $|\mathcal{P}|$ for a given speaker resulting in the similarity matrix

$$\mathbf{S} = [\hat{\mathbf{g}}_1, ..., \hat{\mathbf{g}}_{|\mathcal{P}|}]\,. \tag{3}$$

3. PPG analysis of the TORGO Control Speaker Data

The H-FN PPG model (cf. Section 2.2) is trained on typical speech, and [72] demonstrates generalisation to a number of commonly used typical speech datasets for interpretable acoustic pronunciation distance. In this section, the generalisation of the PPG model to control speech in the *TORGO* is evaluated before analysis of the dysarthric data in Section 4. Therefore, analysis is conducted on (i) differences and agreement between the annotation tokens in the *TORGO* phoneme alignment data and the class tokens $\mathcal{P}$ in the H-FN PPG model (cf. Section 3.1) and (ii) PPG consonant probability and acoustic similarity for the control speaker data (cf. Section 3.2). PPG features could also be used to infer vowel space boundaries (and articulatory working space/perceptual vowel distinction). However, analysis (not shown) of PPG features and vowel space boundaries do not show conclusions beyond analysis of F1 and F2 planar area, and therefore the detailed analysis of this paper will focus on consonants.

3.1. Agreement between TORGO annotation tokens and PPG class tokens

First, we examine and compare the annotation tokens in the *TORGO* alignment data and the class tokens in the H-FN PPG model. *TORGO* has additional annotations for noise and closure tokens relative to the PPG model. The PPG model is trained to output a '<silent>' token (in addition to estimated phoneme tokens), while the *TORGO* annotations include tokens for silence ('sil') and noise ('noi'). For the control speakers, there are only noise annotations in the *TORGO* data for periods of non-speech; on average, the probability P('<silent>'|'noi') (i.e. the probability of '<silent>' in the overall averaged probability distribution for aligned 'noi' annotations) is 96.78%. Furthermore, phoneme aligned non-speech annotation tokens (i.e. all instances of 'sil' and 'noi' aligned posteriorgrams (cf. Section 2.2.2 for a description of the aligned posteriorgram features)) are examined. The PPG model's most likely prediction for non-speech tokens (i.e. arg $\max_{q \in \{sil', noi'\}} P(q|p)$) show a classification accuracy to '<silent>' of 96.93% for the control speakers and 96.51% for dysarthric speakers, showing that the '<silent>' token in the H-FN PPG model has accurately predicted non-speech for both noise and closure tokens in the *TORGO* data.

Another area of interest is the inclusion of closure tokens in the *TORGO* data, denoted 'cl' (e.g. 'pcl' is the closure for the corresponding voiceless bilabial plosive /p/). For control speakers MC01 and MC04, the most likely probability for 'cl' tokens are the corresponding plosive (64.4% on average for P(plosive|cl)), in contrast to the average probability of 21.94% for P('<silent>'|'cl'). For speaker MC02, the highest probability observed for 'dcl', 'gcl' and 'bcl' is '<silent>' (with the corresponding plosive being the second highest probability). On analysis, this is due to the number of silence frames for the closure phase of plosive articulation, and the number of frames from plosive release included in the 'cl' alignment.

The remaining *TORGO* annotation tokens correspond to the remaining PPG tokens (i.e. the ARPAbet [84] tokens in the CMU dictionary). Analysis of acoustic similarity of the consonant tokens for the *TORGO* control data is conducted in the following Section 3.2.

3.2. PPG consonant probability and acoustic similarity for the control data

PPG probability and acoustic similarity for the control speaker data is evaluated in the following. Acoustic similarity is computed on the *TORGO* phoneme aligned data for each control speaker (cf. Section 2.2.3). Figure 2 shows the acoustic phoneme similarity matrix $\mathbf{S}$ as in (3) for TORGO control speakers with PPG tokens on the x-axes and *TORGO* tokens on the y-axes.

Referring to the probability values in Figure 2, the correct probability (i.e. when $q = p$) is the dominant probability for all phonemes. Recent work has demonstrated that relatively higher probability values are observed between similar phonemes in PPG acoustic similarity, such as, between voiced and unvoiced fricatives [72]. Therefore, analysis will be conducted on misclassification probability (i.e. when $q \neq p$) in order to establish patterns in phoneme similarity. While the focus of this work is on dysarthric speech (cf. Section 4), to establish a range of probability values as a basis for comparison first, Table 3 shows the average cumulative values for the 5 highest ranked probabilities in acoustic similarity (i.e. overall averaged distribution $\hat{\mathbf{g}}_p$) across all PPG speech tokens. The first column (Contr.) shows the results for control speakers. Results for dysarthric speakers are shown in the remaining columns and will be analysed in Section 4.

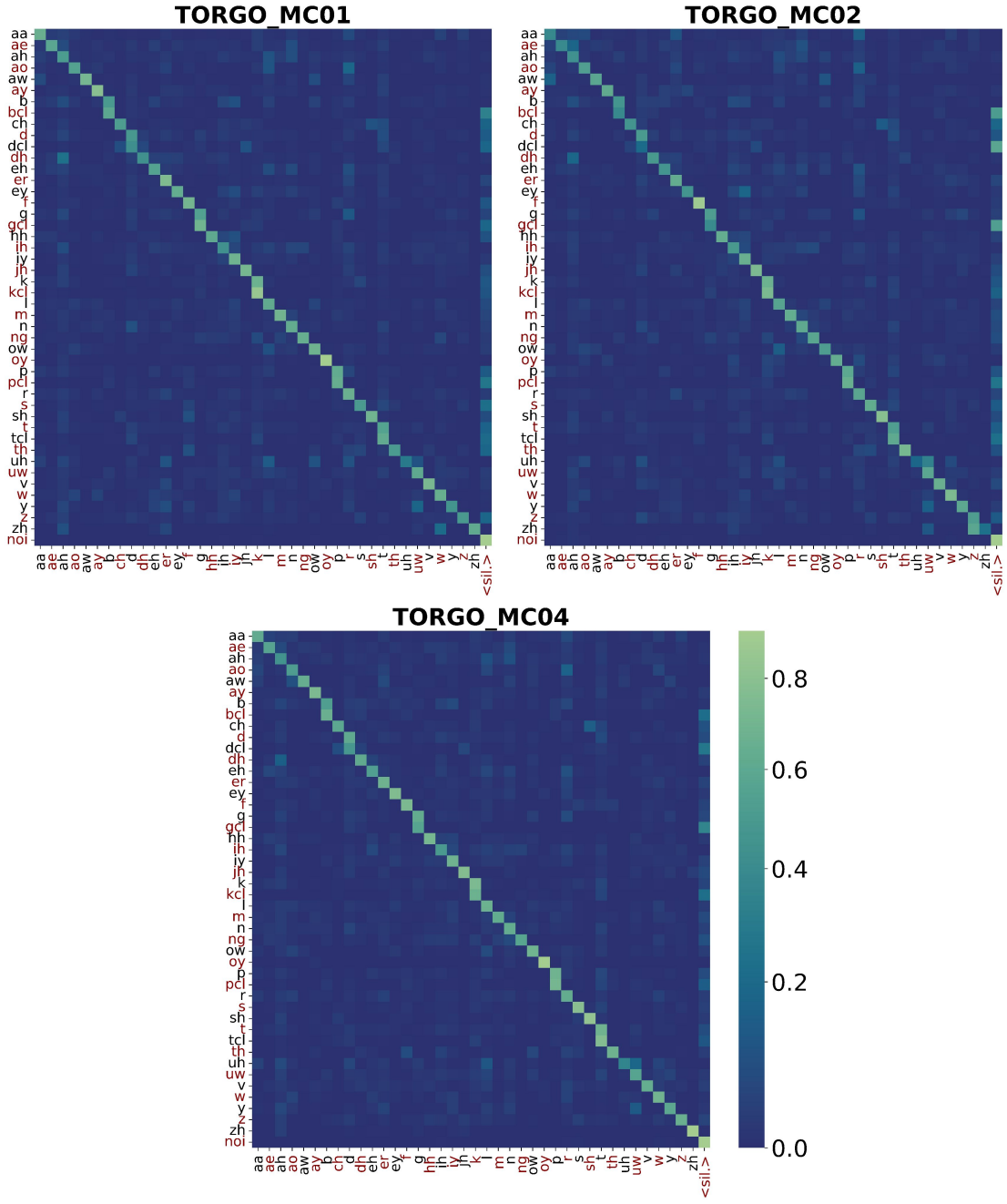

Figure 2: Acoustic phoneme similarity **S** for TORGO control speakers. X-label = PPG tokens, Y-label = TORGO tokens. Values represent probabilities in the range [0, 1].

On average, the most similar phoneme (i.e. Rank2) has a probability value of 9.08% (SD=5.34%). The cumulative probability of the top 4 ranked classes is 79.18%, with relatively low probability increase after the fifth highest value. Finally, it has been noted that factors of phoneme environment, e.g. coarticulation [85] and deletion/insertion errors [62], have been shown to impact on pronunciation models, and the spectral relationships (of e.g. vowel centralisation and assimilation and formant frequencies [86], fricative centroid and peak energy [87]) are well studied. However, the *TORGO* data contains a limited pool of transcripts which are repeated across speakers, and therefore an assumption that the phoneme environment (i.e. contextual variability) is similar across speakers has been taken during analysis unless otherwise stated.

Consonants will be examined under broad categories by manner of articulation. Analysis will focus on acoustic phoneme similarity and misclassification probability between consonants within manner of articulation categories to establish PPG boundary distinction and similarity between phoneme tokens in the control data. Figure 3 shows the acoustic phoneme similarity between plosives (row 1), fricatives (row 2), and affricates, approximants, and lateral approximants, respectively

Table 3: Cumulative PPG probability (%). Comparison of Dysarthric Speakers to Control results (first column, Control speakers).

	Contr.	F01	F03	F04	M01	M04	M05
Rank1 (%)	61.40	26.35	42.92	60.94	26.19	27.53	28.95
SD	11.87	9.66	10.09	11.33	10.46	10.44	11.85
Cumulative sum	61.40	26.35	42.92	60.94	26.19	27.53	28.95
Rank2	9.08	14.47	11.24	9.35	14.82	14.48	13.79
SD	5.34	4.09	4.47	5.71	4.61	4.77	4.40
Cumulative sum	70.48	40.82	54.16	70.29	41.01	42.01	42.74
Rank3	5.13	9.71	7.16	5.41	10.59	9.24	9.03
SD	2.34	2.47	2.40	2.30	3.29	2.83	3.14
Cumulative sum	75.61	50.53	61.32	75.70	51.60	51.25	51.77
Rank4	3.58	7.86	5.40	3.72	8.16	7.03	6.70
SD	1.45	2.10	1.78	1.54	2.51	1.89	2.67
Cumulative sum	79.19	58.39	66.72	79.42	59.76	58.28	58.47
Rank5	2.71	6.26	3.99	3.00	6.40	5.84	5.36
SD	1.20	1.69	1.41	1.04	2.06	1.68	1.68
Cumulative sum	81.90	64.65	70.71	82.42	66.16	64.12	63.83
Entropy	1.23	1.73	1.52	1.26	1.75	1.69	1.65

(row 3). The manner of articulation groups show high average correct probability (plosives=58.05%, fricatives=65.51% and affricates, approximants and lateral approximants=63.94%, 62.70% & 62.10%, respectively), with a large margin relative to the second highest probability. The average correct probability for nasals is 62.58% and an analysis of the nasal consonants is conducted in Section 4.1 (demonstrating acoustic phoneme similarity between nasal consonants with low probability values, and that the similarity between nasal and non-nasal consonants observed is due to the neighboring phonetic environment in control speakers).

Table 4 summarises the predominant trends observed for PPG acoustic similarity for the control speakers. Overall, plosives and fricatives show a pattern of similarity for within manner phonemes with common voicing, and exceptions observed are likely due to devoicing. Although a consistent trend for relatively high probability between voiced and unvoiced plosives is not observed (as in [72]), there are clear plosive PPG boundaries and similarity within voiced and unvoiced plosive classes as expected. Additionally, similarity with low probability values between approximants can be observed, and there is no consistent pattern across control speakers for acoustic similarity between affricates. Acoustic similarity for affricates and fricatives (not shown) was also conducted, with high probability observed for only ʃ|tʃ (on average, 12.32%).

In summary, the analysis conducted in this section shows a high generalisation of the H-FN PPG model to the *TORGO* control data. The correct probability consistently dominates the probability distribution of PPG phoneme tokens, demonstrating distinct phoneme boundaries. Also, relatively higher misclassification probability values are observed for phonemes with a similar manner of articulation, indicating that relation between similar phonemes is captured while distinct phoneme boundaries are maintained in the control data. Analysing patterns of similarity within manner of articulation groups, the results indicate sensitivity to voicing and place (with relatively low probability values observed).

4. TORGO Dysarthric Speaker Data

In this section, PPG features for the dysarthric data will be analysed. Section 3 showed generalisation of the PPG model to the *TORGO* control data, and provided a basis of comparison to the dysarthric data. Table 3 shows the average cumulative values for the 5 highest ranked probability values in acoustic similarity (i.e. overall averaged distribution $\hat{\mathbf{g}}_p$) across all PPG speech tokens. Compared to the control speakers (where PPG identified the dominant correct probability for all phoneme tokens), the highest probability (i.e. Rank1) for dysarthric speakers is lower (by 35.48%, on average) with higher entropy and variance of probability values assigned to the top $2-5$ ranked tokens, indicating higher acoustic phoneme similarity between a higher number of phoneme tokens, and reduced distinction in phoneme boundaries. Therefore, analysis is conducted in the following subsections to evaluate the extent to which differences in correct PPG probability and misclassification (based on acoustic similarity) reflect dysarthric speech processes, with respect to the FDA in speech subscores (cf. *rows in Table 1) and in the context of the variation observed for the Control speakers. Manual listening is also conducted by a SLT where appropriate to confirm that quantitative differences in PPG probability match dysarthric speech processes that are determined by auditory perceptual evaluation. Section 4.1 analyses PPG features and the FDA palate section. Section 4.2 introduces dysarthria and articulation impairment, and the tongue and lip FDA aspects are analysed in Section 4.3 and Section 4.4, respectively. Section 4.5 and Section 4.6 analyse laryngeal and intelligibility FDA aspects, respectively.

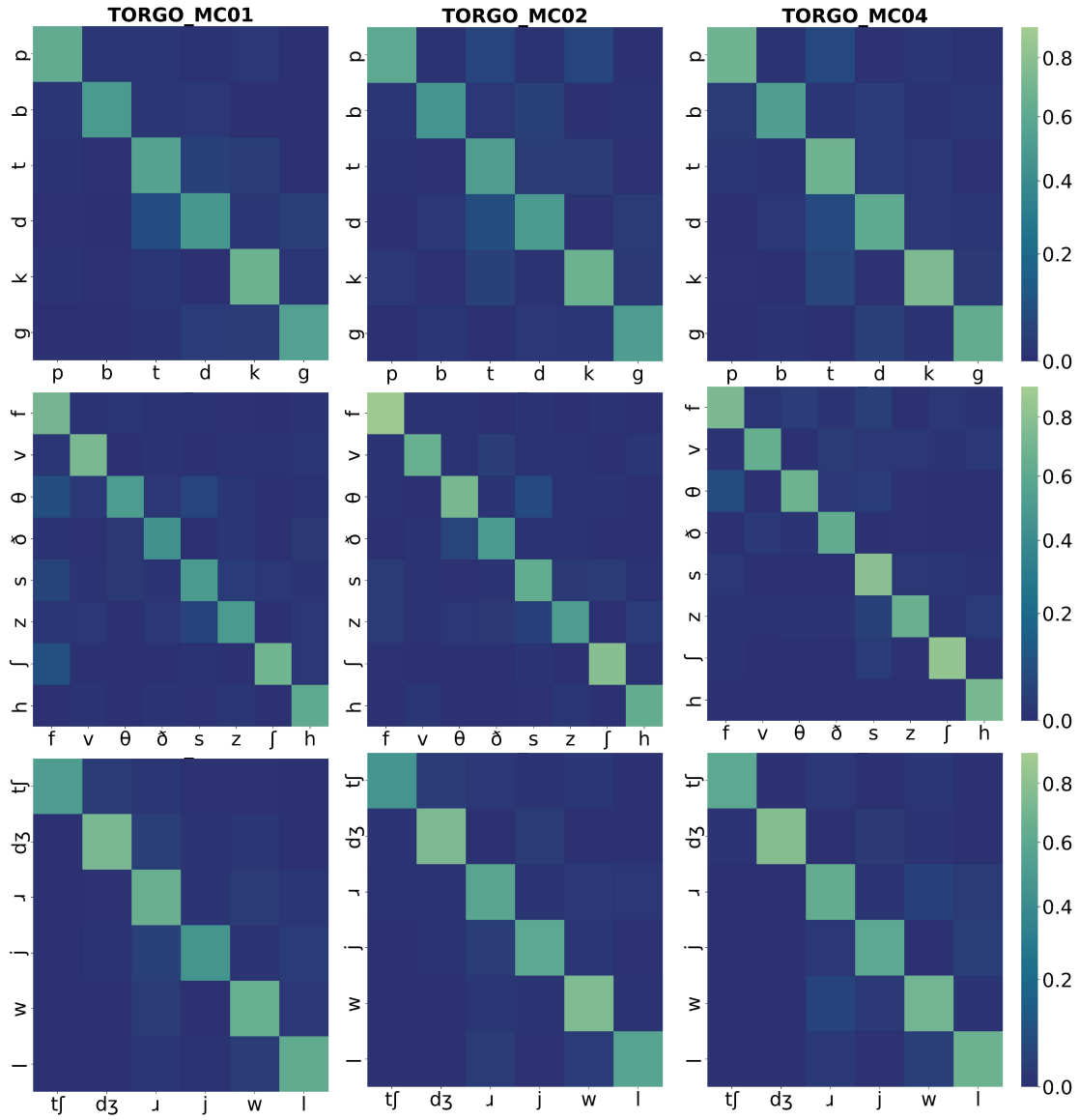

Figure 3: Control speaker acoustic similarity **S** for plosives (row 1), fricatives (row 2), and affricates, approximants & lateral approximants (row 3).

4.1. FDA Soft Palate rating

In this section, analysis is conducted on PPG features with respect to the FDA soft palate ratings (cf. Table 1, row 3). Dysarthria secondary to CP and ALS can cause reduced range and force of velopharyngeal movements [77], and inadequate velopharyngeal closure can result in hypernasality, and weak and imprecise consonants [75]. The soft palate in speech FDA subscore is rated on the perceptual evaluation of nasal resonance and nasal emission during spontaneous conversation. Therefore, analysis will focus on PPG acoustic similarity to nasal tokens (i.e. /m, n, ng/) and whether the presence and severity of deviant nasal resonance and nasal emission is captured. However, research has demonstrated both coarticulatory effects of nasal consonants [91] and particular velar height for vowels (generally higher for close vowels, although the velopharyngeal port is still closed) [92]. Thus, analysis is conducted under two conditions: (i) on all phoneme aligned data (i.e. including adjacent nasal tokens, denoted +Adj. nasal) and (ii) on phoneme tokens with no adjacent nasal tokens (−Adj. nasal).

Figure 4 shows acoustic similarity to nasal phonemes for MC01 (control) and F03 & F04 (dysarthric speakers with soft palate rating 'A' (normal function)). For MC01, nasalised vowels (with relatively low probability values) can be observed, which are not observed without adjacent nasal tokens (right side). The same pattern is also observed with other control speakers. For speaker F04, hypernasalised vowels are significantly reduced without adjacent nasal tokens, whilst hypernasality with plosives remain, indicating capture of nasalised plosives irrespective of nasal coarticulation (and incomplete palate closure during plosive release). Similarly, for F03, the degree of hypernasalied vowels and plosives are reduced with-

Table 4: Predominant trends observed in acoustic similarity **S** for Control speakers.

Plosives	
As highlighted in the analysis in Section 3.1, there is similarity between plosives and the '<silent>' token, which corresponds to silence frames and the closure phase of plosive articulation.	For unvoiced plosives, there is on average 10.45% similarity to the '<silent>' token, and '<silent>' is consistently the second highest probability. For voiced plosives, the average similarity is 4.99%.
The highest similarity observed between plosives is P(t|d)=7.30%. Common devoicing of /d/ in the English language (e.g. final devoicing [88] and voicing assimilation [89]) may account for the observation.	In support, there is a lower similarity between /t/ & /d/ (on average, 2.65% probability).
Beyond this exception, there is a general pattern for similarity between plosives with common voicing with low probability values.	For unvoiced plosives, the highest average similarity for control speakers is observed for p|t (3.85%), k|t (2.87%), and p|k (2.26%), for voiced plosives, b|d (2.35%), g|d (2.23%), and d|g (1.94%).
Fricatives	
A similar trend to plosives can be observed. An exception to the overall pattern is low probability values between the voiced and unvoiced alveolar fricatives (on average, 3.69% for s|z). Devoicing of sibilants may account for this observation [90], and in support there is lower similarity for z|s (on average, 1.63%).	Generally, there is an overall pattern of similarity with low probability values between unvoiced fricatives, with the highest values observed for the unvoiced dental fricative (on average, 5.08% for f|θ and 4.53% for s|θ), and unvoiced sibilants to /f/ (on average 3.01% for f|ʃ, and 2.3% for f|s).
Affricates, approximants, and lateral approximants	
Similarity with low probability values between approximants can be observed	Highest values observed are between the alveolar and labial-velar approximant (on average, 2.29% for ɹ|w and 2.30% for w|ɹ), and the palatal and alveolar approximant (on average, 1.97% for ɹ|j).

out nasal coarticulation. Further data and analysis would be required to establish whether non auditory-perceptible hypernasality is present in 'A' rated speakers vs. controls. Informal evaluation by an SLT concluded a mild nasal speech quality, otherwise nasality was deemed to be within normal limits. However, the important trend is that for both F04 and F03, acoustic similarity to nasal tokens with low probability values for adjacent nasal tokens is observed, which are reduced without adjacent nasal tokens.

The dysarthric speakers with palate impairment are shown in Figure 5. Acoustic similarity to nasal tokens is maintained in phonemes without adjacent nasal tokens (although there is marginal reduction in probability values), indicating that hypernasality has been captured irrespective of nasal coarticulation. Therefore, speakers with and without soft palate impairment are easily distinguished. Hypernasality is captured with vowels, plosives, and affricates which is consistent with research on velopharyngeal dysfunction [93, 94]. Furthermore, there is no consistent pattern for similarity to nasal tokens for periods of non-speech (i.e. aligned silence and noise), indicating that the distinctions are not based on the acoustic recording environment. To examine if the severity of deviant nasality can be distinguished, Kolmogorov-Smirnov (KS) tests are conducted on phoneme aligned token ($\bar{\mathbf{g}}_p$) misclassification probability to /n/ (i.e., $P(q = /\text{n}/)$). Table 5 shows the results for pairwise KS test for soft palate groups by manner of articulation.

Statistically significant differences in the distribution of misclassification probability for all manner of articulation groups were observed between Control and 'A' rated dysarthric speakers ($P < 0.001$), with the largest KS statistic observed for af-

Table 5: Pairwise soft palate group Kolmogorov-Smirnov test results by manner of articulation (phoneme aligned token ($\bar{\mathbf{g}}_p$) misclassification probability values to /n/).

Manner	KS Stat (D)	P-value	Sig.
Comparison: Control vs. A			
Affricate	0.4176	< 0.001	***
Approximant	0.3186	< 0.001	***
Diphthong	0.2936	< 0.001	***
Fricative	0.2411	< 0.001	***
Lateral Approximant	0.3245	< 0.001	***
Monophthong	0.2534	< 0.001	***
Nasal	0.0980	< 0.0001	***
Plosive	0.2956	< 0.001	***
Comparison: A vs. B			
Affricate	0.1786	0.5761	
Approximant	0.2163	< 0.001	***
Diphthong	0.1833	< 0.001	***
Fricative	0.1634	< 0.001	***
Lateral Approximant	0.2421	< 0.001	***
Monophthong	0.0771	< 0.001	***
Nasal	0.0882	0.0249	*
Plosive	0.0564	0.0373	*
Comparison: B vs. D			
Affricate	0.2857	0.4393	
Approximant	0.1124	0.1911	
Diphthong	0.1519	0.1890	
Fricative	0.2495	< 0.001	***
Lateral Approximant	0.2607	0.0085	**
Monophthong	0.1424	< 0.001	***
Nasal	0.2422	< 0.0001	***
Plosive	0.3144	< 0.001	***

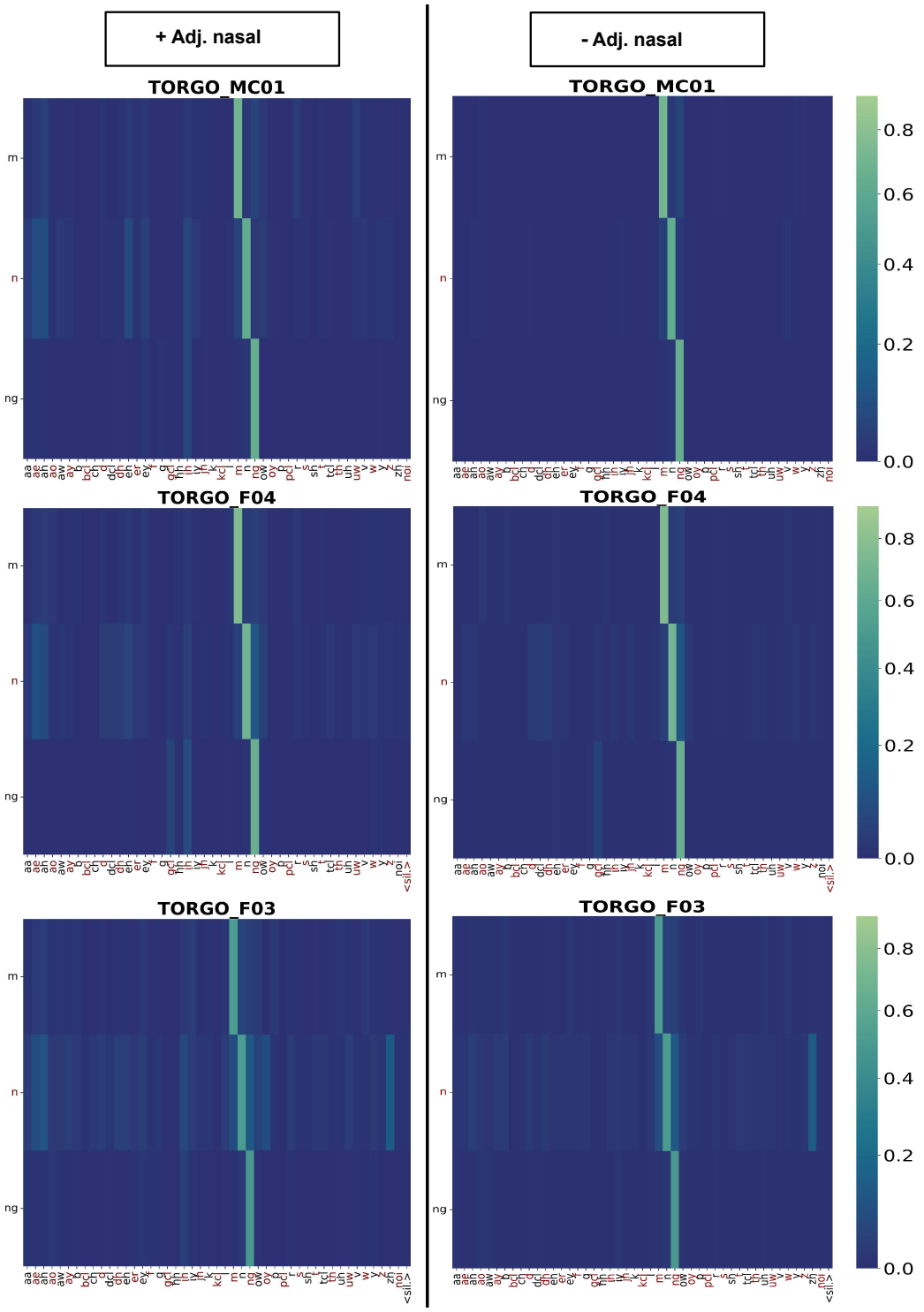

Figure 4: Acoustic similarity to nasal tokens (i.e. /m, n, ng/) for control speaker MC01 and dysarthric speakers F03 & F04 (no palate impairment). The left column shows acoustic similarity for the nasal tokens calculated on all phoneme tokens, and the right column is calculated only on phonemes with no adjacent nasal tokens.

fricates, approximants, plosives and vowels (diphthongs and monophthongs, respectively). For 'A' vs. 'B' speakers all manner of articulation groups apart from affricates are statistically different ($P < 0.0373$), with the largest relative differences observed for approximants, diphthongs and fricatives. For 'B' vs. 'D' speakers, statistically significant differences ($P < 0.0085$) were observed for plosives, lateral approximants, fricatives, nasals and monophthongs.

Finally, speaker F01 exhibits significant nasal emission and hypernasality of consonants in audio samples (and this is doc-

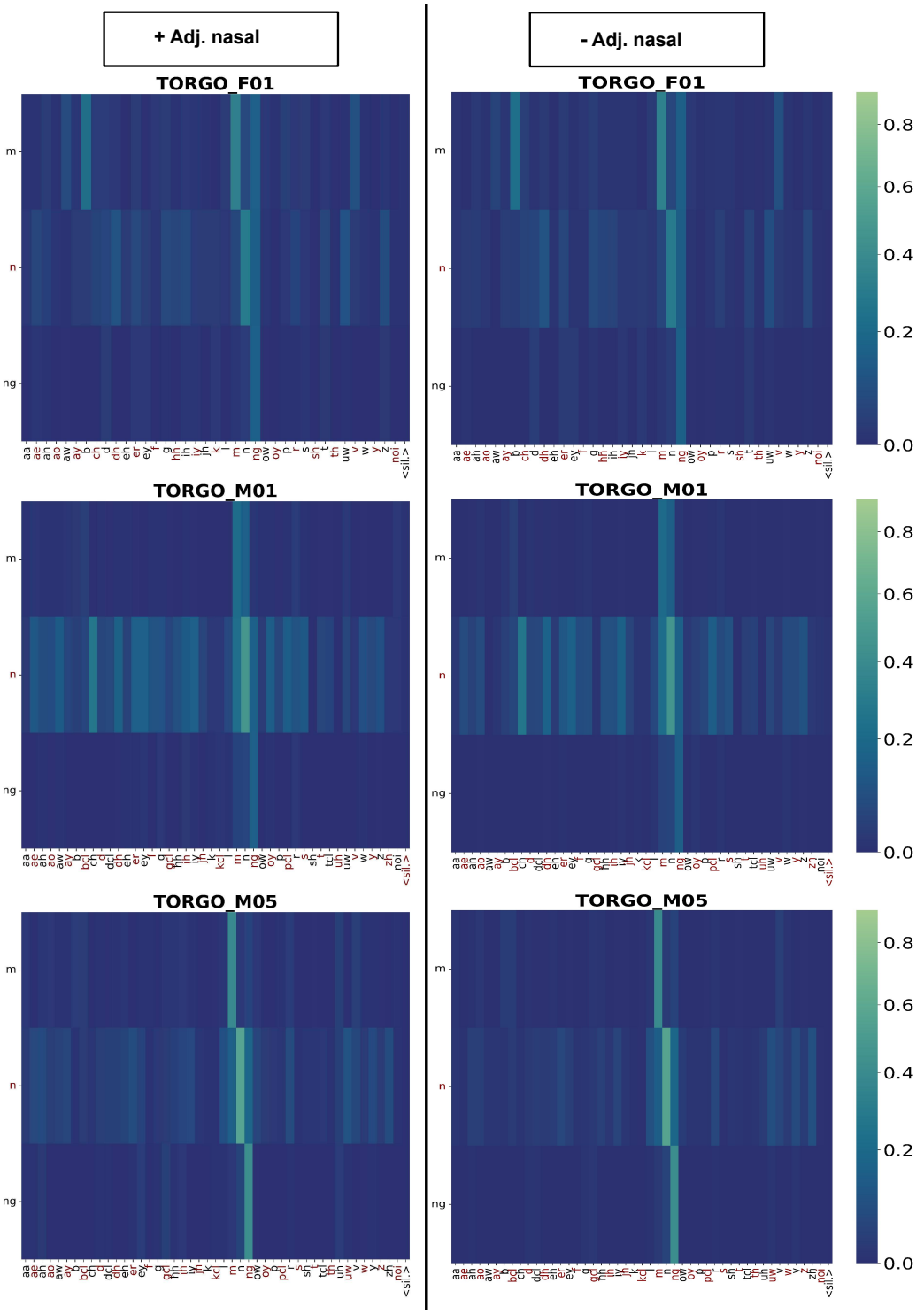

Figure 5: Acoustic similarity to nasal tokens (i.e. /m, n, ng/) for dysarthric speakers with soft palate impairment. Soft palate subscore: F01=D, M01=B, M05=B.

umented in the FDA assessment [11]). Higher misclassification probability to /m/ relative to other dysarthric speakers can be observed (e.g. P(m|b)=23.10%, P(m|ng)=14.64%, and P(m|v)=8.44%). The other dysarthric speakers do not exhibit nasal emission and have low misclassification probability values to /m/ (cf. 'm' columns, Figure 5).

In summary, the presence of deviant nasality can be interpreted from PPG misclassification probability. Furthermore, PPG probability values between severity groups are statistically different, showing that severity of palate impairment in *TORGO* can be distinguished. Misclassification probability to /m/, i.e. P(q=m) distinguishes the presence of nasal emission for speaker F01, however there are no other speakers with this feature to analyse severity and nasal emission.

4.2. Dysarthria and tongue & lip impairment

In the following sections, acoustic similarity and tongue and lip FDA aspects are examined. Imprecise consonants are attributed to reduced range and force of articulatory movement in spastic & flaccid dysarthria, and imprecise consonants and irregular articulatory breakdown are attributed to dysrhythmia (disturbed rhythm) and inaccurate direction of repetitive articulatory movement in ataxic dysarthria [77]. Due to the broad and complex scope of articulatory impairment, the analysis is limited to the presence of lingual and labial dysarthric processes (i.e. the severity of these processes is not analysed). The FDA in speech jaw subscore is rated on the observed position of the jaw during conversational speech. Although jaw movements have been shown to e.g. affect vowel production [95] and impact on tongue movement [96] in dysarthria, it is difficult to delineate jaw processes from other articulatory parameters from PPG class probability. It would be possible to evaluate jaw electromagnetic articulography (EMA) sensor data and articulation. However, conclusions drawn would be reliant on the EMA data (and not interpretation of PPG information). Furthermore, the jaw section was removed from the FDA-2 as jaw information did not assist diagnosis [34] (cf. Section 2.1.1). Therefore, jaw articulation is not analysed for the scope of this paper.

Finally, the number of PPG phoneme frames (i.e. duration) can also be analysed. [11]'s evaluation includes analysis of duration, and therefore this is not repeated but the results would provide supplementary information on the speed and consistency of articulation.

4.3. FDA Tongue rating

The FDA in speech tongue subscore is rated on tongue movement in conversation, including observation of the accuracy, speed, and labour of movement, and perception of phoneme distortion and consonant omission. In the following, plosives, fricatives, approximants, and lateral approximants with active tongue articulation are examined in context of lingual impairment and PPG probability.

4.3.1. Plosives and FDA tongue rating

Plosives require complete closure between two articulators to obstruct airflow, and subsequent separation of the articulators to release the compressed air. Imprecise plosives due to e.g. weakness, incoordination or reduced range of motion include distortions or substitution based on variability in manner (e.g. incomplete closure [97]), place of articulation [98], and premature plosive release/inadequate air compression may lead to weak plosive sound [99]. Inadequate velopharyngeal closure (i.e. soft palate impairment) may lead to air escape through the nasal cavity, causing nasalised plosives, difficulty creating compressed air, and nasal emission (cf. Section 4.1 for an analysis of deviant nasality). Impaired laryngeal timing or control can result in reduction or loss of the voice/voiceless distinction [100], and this will be explored further in Section 4.5.

Acoustic similarity between plosives for dysarthric speakers is shown in Figure 6. The control data shows distinction between plosive tokens, and a general trend of similarity with low probability values within voiced and unvoiced plosive classes (cf. Section 3.2). In order to evaluate plosives and lingual impairment, the plosives with active lingual articulation (i.e. alveolar and velar plosives) are examined. The acoustic similarity profiles for F04 and F03 ('A' in speech tongue rating) are similar to the control speaker data, with a comparable range of misclassification probability values.

Table 6 summarises the predominant trends observed for PPG acoustic similarity for the dysarthric speakers with rated FDA tongue impairment. The trends show overall higher similarity relative to control speakers between plosives of neighbouring place of articulation indicating reduced distinction in lingual plosive boundaries, and the 'silent' token which may indicate weak plosives. In summary, the acoustic similarity profile (and phoneme distinction with neighbouring plosive phonemes) differentiate speakers with and without FDA tongue impairment, although reduction in voicing contrast is the dominant process captured by PPGs. Misclassification probability to neighboring plosive phonemes correspond to dysarthric processes on manual evaluation, e.g. fronting and imprecise lingual place of closure. Higher misclassification probability may indicate weak plosive articulation, but this is not straightforward to interpret.

4.3.2. Fricatives and FDA Tongue rating

Fricatives require fine motor control to maintain a narrow channel (close approximation) and turbulent airflow between two articulators. If these are not maintained, this will lead to distortion or substitution of the fricative [101] (and changes in spectral characteristics [102]). Figure 7 shows acoustic similarity between fricatives and plosives, approximants and lateral approximants for MC01 and 'A' (normal function) rated dysarthric speakers, and Figure 8 shows the acoustic similarity for dysarthric speakers with in speech lip and tongue impairment (i.e. rated 'A/B' or higher).

The fricatives with active lingual articulation (i.e. dental, alveolar, and post-alveolar fricatives) are examined, and Table 7 summarises the predominant trends observed for PPG acoustic similarity. For control speakers, low similarity values overall are observed. F04 shows a similar profile, and F03 shows a marginal reduction in phoneme distinction. For dysarthric speakers with rated tongue impairment, higher similarity values indicate distortion due to complete closure and open approximation of the narrow channel required for frication. In summary, speakers with and without FDA tongue impairment are distinguished. Higher probability values observed in acoustic similarity for fricatives to plosives and approximants (which correspond to dysarthric processes on manual evaluation) indicate that PPGs have captured phonological processes related to fine motor control and maintaining a narrow opening and turbulent airflow between the tongue and passive articulator (i.e. narrow channel to complete closure for fricative stopping, and narrow channel to fricative approximantisation).

4.3.3. Approximants and FDA tongue rating

Approximants require open approximation (i.e. a narrower approximation than vowels, but wider than for fricatives in

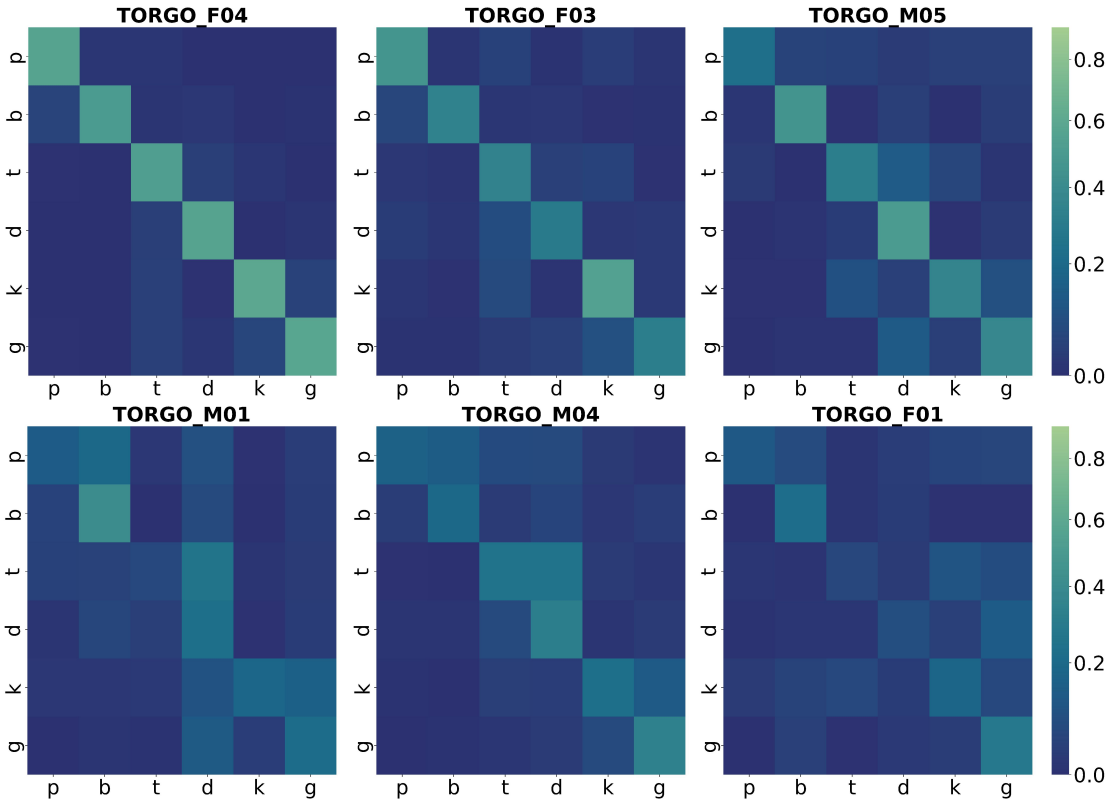

Figure 6: Acoustic similarity for plosives (Dysarthric speakers).

Table 6: Predominant trends observed for Dysarthric speakers with tongue impairment

Alveolar and velar plosives	
Higher similarity between plosives of neighbouring place of articulation overall relative to control speakers, indicating reduced distinction in lingual plosive boundaries. However, it can be observed that reduction in voicing contrast is the dominant process captured in unvoiced alveolar plosives (cf. Section 4.5 for further analysis).	For example, higher acoustic similarity can be observed for P(d|g) for M05 & M04 (by 10.17% and 10.41% relative to the control speaker average, respectively), indicating fronting of the voiced velar plosive towards the alveolar region (which can be heard in audio samples, e.g. speaker M05, session 1, array mic wav 24 (M05, S1, AM24)). Marginally higher similarity between unvoiced alveolar and velar plosives are observed for speakers M05 and F01 (by 5.61% and 4.36%, respectively, relative to control speakers).
Increased acoustic similarity to the 'silent' token for plosives (not shown). However, analysis in Section 3.1 demonstrated correlation between the number of non-speech frames and P('silent'|plosive) in the control data and therefore may not be a reliable indicator of weak plosives unless this variable is considered.	Dysarthric speakers with tongue impairment have 10.88% higher acoustic similarity on average between alveolar and velar plosives and 'silent' in comparison to control speakers, and the probability values distinguish 'C/D' and 'D' rated speakers by 5.11% on average.

order to not cause turbulent airflow) between two articulators [101]. Approximants can also require precise shaping of articulators, e.g. /r/ requires retroflex (curled back) apical (tongue tip) open approximation to the alveolar region, and /w/ requires a rounded lip posture and raising of the tongue dorsum (back) toward the palatal region. Figure 9 shows the acoustic similarity for approximants and lateral approximants to approximants, lateral approximants, plosives and close to open-mid monophthongs for control speaker MC01, and the dysarthric speakers with 'A' tongue and lip ratings. Figure 10 shows the acoustic similarity for the dysarthric speakers with rated in speech tongue and lip impairment.

Alveolar and palatal approximants, and the alveolar lateral approximant (i.e. approximants and lateral approximants with active tongue articulation) are examined. For the control and 'A' rated speakers, there are low probability values for similarity to plosives, and (other, as applicable) approximants and lateral approximants (0.75% and 1.45% on average, respectively). For dysarthric speakers with tongue impairment, distinction between approximants and lateral approximants is generally maintained. There is marginally higher similarity for P(ɹ|l) (by 2.63%) and P(l|ɹ) (by 1.9%) on average relative to control speakers, otherwise the misclassification probability values are comparable. Speaker F01 exhibited the highest similarity for

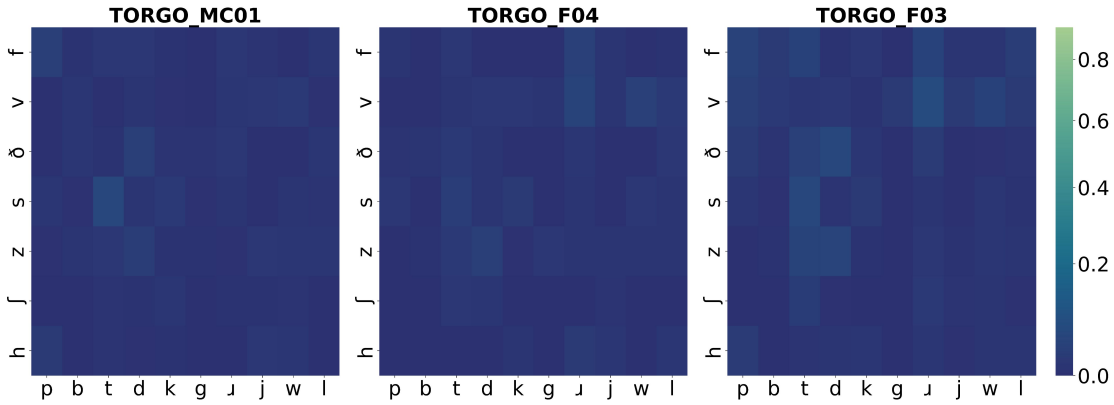

Figure 7: Acoustic similarity for fricatives to plosives, approximants, and lateral approximants (MC01 = Control, F04 & F03 = 'A' tongue and lip rating).

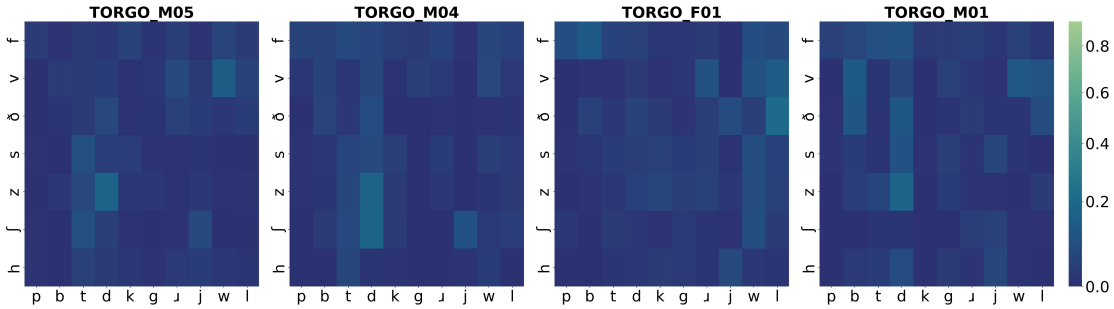

Figure 8: Acoustic similarity for fricatives to plosives, approximants, and lateral approximants (dysarthric speakers with rated in speech tongue and lip impairment).

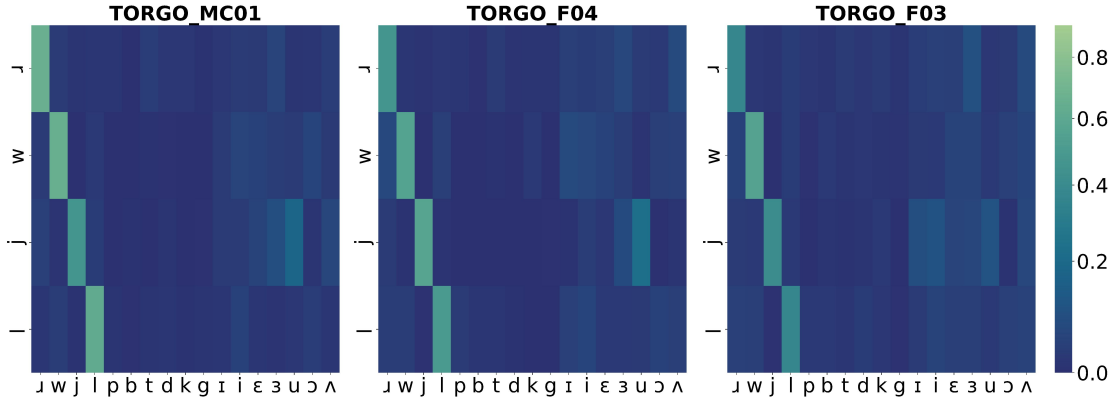

Figure 9: Acoustic similarity for approximants and lateral approximants to approximants and lateral approximants, alveolar and velar plosives, and close to open-mid monophthongs (control and 'A' rated speakers).

P(l|ɹ) (6.24%), and reduced retroflexion can be heard in audio samples (e.g. F01, S1, AM37). In summary, differences in PPG probability for distortion and substitution of lingual approximants were not clear cut. Analysis of similarity to plosives, fricatives and monophthongs (open approximation vs. complete closure, narrow approximation, and open vocal tract, respectively) was also conducted, but no significant trends were observed.

4.4. FDA Lip rating

The FDA in speech lip subscore is rated on the observed movement of the lips in conversation. In the following, plosives, fricatives, and approximants with active lip articulation are examined in the context of labial articulation and PPG probability. The lip subscore is rated on the movement/shape of the lips for all sounds (and not just bilabial and labiodental consonants). Therefore, lip articulation and vowels were also evaluated (not shown), but patterns between lip impairment and e.g. lip rounding and vowel articulation were not captured in PPGs. Analysis of similarity to consonants with high sonority and labial articulation was also conducted, and the results demonstrated similarity with rounded vowels and the labial-velar approximant, but no deviation in pattern for dysarthric speakers with rated lip impairment was observed, and therefore this analysis is not shown.

Table 7: Predominant trends observed for fricatives and tongue impairment

Dental, alveolar and post-alveolar fricatives	
For control speakers, low similarity values overall can be observed between lingual fricatives and lingual plosives (and an average similarity between lingual fricatives and approximants & lateral approximants of 0.50%). Relatively higher values are observed likely due to the phonetic environment, e.g. consonant clusters.	e.g. For MC01, P(t|s)=4.94% and P(k|s)=1.40%. Acoustic similarity was calculated on the /s/ phoneme token without neighbouring lingual plosive tokens, and similarity was reduced (by 1.7% and 0.87%, respectively), indicating at least partial contribution from phonetic environment (e.g. consonant clusters).
For 'A' rated dysarthric speakers, F04's probability value ranges and profile are comparable to the control data, and F03 shows marginally increased values, indicating a relatively mild reduction in phoneme distinction.	For F04, marginally increased values relative to control speakers between fricatives and alveolar plosives (by 1.76% and 1.05% on average for /t/ and /d/, respectively), and the alveolar approximant (by 0.98% for /r/).
For dysarthric speakers with rated tongue impairment, higher similarity to plosives, approximants and lateral approximants overall can be observed, indicating distortion due to complete closure and open approximation of the narrow channel required for frication.	For M05, M04, and M01, on average P(/d/|/z/) is 13.44% higher relative to control speakers. On manual evaluation, complete closure to the alveolar region can be heard (e.g. M04, S1, AM12). Speakers M05 & M04 also show higher similarity for P(/j/|/ʃ/) (9.55% and 6.72%, respectively), and open approximation to the palatal region and reduced airflow can be heard during (pre- and inter-vocalic) fricative articulation (e.g. M05, S2, HM132). For F01, some distortion can be perceived in fricatives on manual evaluation, but it can be heard that turbulent airflow (and therefore close approximation) is maintained in audio samples.
For speakers with both soft palate and tongue impairment (F01 and M01), there is a high similarity for fricatives to the alveolar nasal, indicating hypernasalised fricative stopping.	e.g. on average P(n|ð)=12.62% and P(n|z)=8.22%. However, analysis in Section 4.1 demonstrates similarity to /n/ due to deviant nasality, independent of phonological stopping, and therefore it is difficult to delineate these processes.
For the voiced dental fricative /ð/, all control speakers and F04 show low probability values for similarity to the voiced alveolar plosive /d/ (1.32% on average). F03 has relatively higher probability values. The dysarthric speakers with tongue impairment show higher probability values. It is not straightforward to differentiate potential TH-stopping in e.g. accent variation in Canadian English [103], and stopping/reduced frication and phonological backing & simplification of /ð/ to /d/.	F03 shows relatively higher probability (by 3.93%), and reduced frication/stopping can be heard in audio samples (e.g. F03, S1, AM30). M01 shows increased similarity (by 9.78% on average compared to control speakers & F04), and significant fricative stopping can be heard (e.g. M05, S1, AM11). Increased similarity to the alveolar lateral approximant can also be observed, particularly for F01 (by 18.76% relative to the control average). However, on auditory evaluation, primarily backing and stopping can be heard in F01's articulation.

4.4.1. Plosives and FDA lip rating

Examining misclassification probability between plosives and phonemes with lip rounding (not shown), marginally higher similarity values between the voiced bilabial plosive /b/ and the voiced labial-velar approximant /w/ for dysarthric speakers with lip impairment is observed, but this was not significant. Therefore, errant lip processes in plosives are not captured by PPG information in the *TORGO* data, and analysis of other FDA aspects shows that bilabial plosive misclassification predominantly captures deviant nasality (cf. Section 4.1) (and potentially weak plosive release (cf. Section 4.3.1)) and errant voicing (cf. Section 4.5).

4.4.2. Fricatives and FDA lip rating

Fricatives in Figure 7 and Figure 8 are examined in this section. Labiodental fricatives require close approximation between the lower lip and upper teeth. The control speakers show distinction between the voiced labiodental fricative /v/ and voice labial-velar approximant /w/ (with 2.09% acoustic similarity on average). F04 and F03 ('A' lip rating) show marginally higher values (by 1.14% and 1.41%, respectively). Additionally, for the voiced dental fricative /ð/, the control speakers and dysarthric speakers with no rated lip impairment show low similarity values to /b/. Table 8 summarises the predominant trends for dysarthric speakers with lip impairment. In summary, higher probability values observed in acoustic similarity for labiodental fricatives and approximants (which correspond to dysarthric processes on manual evaluation) indicate that PPG features have captured phonological processes related to motor control, lip posture and labiodental contact & frication.

4.4.3. Approximants and FDA lip rating

Examining Figure 9 and Figure 10, there is a higher similarity between the alveolar approximant and voiced labial-velar approximant for dysarthric speakers with lip impairment. M05 and F01 show the highest values for P(w|r) (8.65% and 5.23%, respectively), and lip rounding (with reduced apical retroflexion) can be perceived during articulation (e.g. M05, S1, AM46 and F01, S1, AM37). No further trends are observed.

4.5. FDA Laryngeal rating

The FDA in speech laryngeal subscores are rated on perceptual evaluation of phonation and appropriate volume and pitch in conversational speech. The PPG token categories do not directly encode information regarding voice quality and characteristics, e.g. pitch, stress, harshness, and loudness. Loudness and weak plosive articulation may be represented in consonant precision, e.g. increased probability of 'silence' with

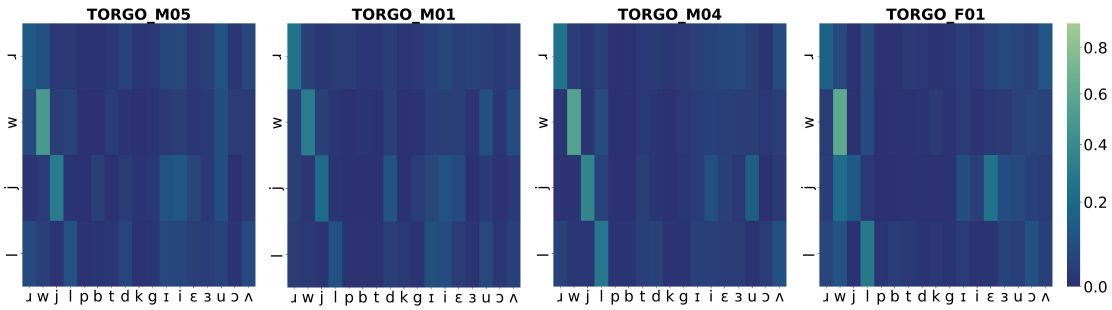

Figure 10: Acoustic similarity for approximants and lateral approximants to approximants and lateral approximants, alveolar and velar plosives, and close to open-mid monophthongs (in speech lip and tongue impairment).

Table 8: Predominant trends observed for Dysarthric speakers with lip impairment

Fricatives and lip impairment	
Higher probability values for P(/w/|/v/), indicating capture of labialisation/approximant substitution. The phonological process also involves movement of the tongue dorsum towards the velum, and it can be observed that dysarthric speakers with rated lip impairment also had rated tongue impairment (cf. Table 1). Dysarthric speakers also had higher probability values that indicate reduced labiodental frication and reduced lip rounding.	In particular, speakers with 'C/D' rated lip impairment had higher values for P(/w/|/v/) (i.e. M01 and M05, by 9.55 and 11.60%, respectively), indicating capture of labialisation/approximant substitution (e.g. M05, S1, AM61). F01 and M05 have higher values between /v/ and /l/ & /r/ (on average, 7.60% for r|v and 8.68% for l|v). In audio samples, reduced labiodental frication with alveolar approximation and a variable reduced degree of lip rounding can be heard in articulation (e.g. F01, S1, AM20). In instances of multisyllabic words with a neighbouring syllable with alveolar articulation, alveolar approximation with no labiodental frication can be heard for speaker M05 (e.g. M05, S2, HM102).
The dysarthric speakers with lip impairment show higher probability values for P(/b/|/ð/). M01 shows the highest similarity (by 10.45% relative to control speakers).	It cannot be clearly discerned from manual evaluation if there is bilabial or linguolabial contact (particularly with the absence of visual information) during articulation (e.g. M01, S1, AM56).

plosives as shown in Section 4.3.1, but this is difficult to disentangle. Research has demonstrated that acoustic measures can effectively capture voice parameters (e.g. time-based acoustic measures and perturbation analysis for Type I signals (i.e. approaching periodic in nature) [104], spectral analysis for Type II and III signals (i.e. with modulations/subharmonics, and approaching random aperiodic vibration, respectively) [105]), and multi-dimensional analysis of dysarthric voice characteristics [100]. However, PPGs may potentially capture errant voicing processes. [11] previously evaluated the substitution errors caused by errant voicing in the *TORGO*, and the analysis demonstrates that dysarthric and non-dysarthric speakers can be differentiated. Some substitution errors are annotated in the phoneme aligned data, and therefore the analysis for this section will focus on the voiceless/voicing phoneme distinction in PPGs, which has adjunctive potential to acoustic measures (in combination with an evaluation of substitution errors) for a full profile of voice dimensions.

Analysis of control speaker data in Section 3.2 shows that acoustic similarity with low probability values is observed within voiced and unvoiced plosive classes, with the exception of t|d (with low probability values observed for d|t). Therefore, acoustic similarity between an unvoiced plosive and a voiced plosive with the same place of articulation is not established, indicating a clear PPG distinction for unvoiced plosives. Acoustic similarity for plosives for dysarthric speakers is shown in Figure 6. The acoustic similarity profile for F04 and F03 ('A' in speech laryngeal rating) is similar to the control speaker data, with a comparable range of misclassification probability values.

For the dysarthric speakers with rated laryngeal impairment, there is significantly higher similarity between unvoiced and voiced plosive pair misclassification (on average 18.48%, 11.73%, and 10.79% higher than control speakers for P(d|t), P(b|p), and P(g|k), respectively), indicating reduced distinction in voicing contrast (and potentially voice onset time (VOT) that distinguishes voiced and voiceless plosives [106]). F01 shows some deviation from this pattern, but this speaker's plosive articulation is characterised by weak plosive release with hypernasality & nasal emission (cf. Section 4.2 and Section 4.1). Research has demonstrated shorter VOT (and stop gap duration) in a number of dysarthria types (for both hypothesised underlying mechanisms of hypertonicity of the larynx [107] and dysrhythmia [108], which in turn are influenced by factors of speech rate, and lung volume at speech initiation [109]).

4.6. FDA Intelligibility rating

The previous sections on FDA aspects have highlighted the dysarthric processes that contribute to speech intelligibility impairment. For clinical dysarthria assessment, it is important to characterise the dysarthric features (and therefore dysarthria profile) underlying changes in phoneme distinction that contribute to intelligibility impairment. Nevertheless, the relationship between PPG probability and speech intelligibility is briefly examined. The FDA sentence intelligibility subscore is

rated on accuracy in a sentence interpretation task. Studies have shown correlation between phoneme & word error rate scores and speech intelligibility for dysarthric speech [110], although error types are often not directly interpretable.

Table 9: Phoneme token accuracy by correct probability in acoustic phoneme similarity (in %).

	Control	F04	F03	M05	M04	F01	M01
Intelligibility	A	A	A	C	D/E	D/E	D/E
Vowels	100	100	100	71.43	85.71	64.29	57.14
Consonants	100	100	100	68.18	59.09	31.82	40.91
Average	100	100	100	69.44	69.44	44.44	44.44

Table 9 shows phoneme token accuracy, where accuracy is calculated as the number of instances where the correct probability (i.e. when q=p) is the dominant probability in acoustic phoneme similarity ($\hat{\mathbf{g}}_p$), divided by the total number of tokens. Control and dysarthric speakers rated 'A' for intelligibility had 100% accuracy for vowel and phoneme tokens analysed, differentiating speakers in the *TORGO* dataset with and without rated FDA intelligibility impairment. Examining average accuracy ('Average' row in Table 9), M05 and M04 have the same average accuracy and therefore 'C' and 'D/E' speakers are not identified by the average probability alone, but the accuracy for vowels and consonants differs. The consonant accuracy score ('Consonants' row) does differentiate severity of FDA intelligibility rating, although there is a relatively wide range of values for 'D/E' speakers, and there are no 'A/B'-'B/C' rated speakers to compare. A Kruskall-Wallis (KW) test shows significant differences in acoustic similarity ($\hat{\mathbf{g}}_p$) correct probability values across intelligibility severity groups (H-statistic = 104.30, p < 0.0001). Table 10 shows the adjusted P-values for a pairwise post-hoc comparison (Dunn's test with Bonferroni correction) between intelligibility severity groups. There is no significant difference between the Control and 'A' group (P=0.2003), and significant differences are observed between the control group & 'C' and 'D/E' groups, and 'A' & 'C' and 'D/E' groups (P < 0.0001). However, there is no significant difference between 'C' and 'D/E' groups.

Table 10: Post-hoc pair-wise comparison between intelligibility severity groups (acoustic similarity ($\hat{\mathbf{g}}_p$) correct probability values).

Comparison	Adjusted P-value
Control vs. A	0.2003
Control vs. C	< 0.0001
Control vs. D	< 0.0001
A vs. C	< 0.0001
A vs. D	< 0.0001
C vs. D/E	0.9516

Therefore, correct probability in acoustic phoneme similarity shows some trends but lacks sensitivity in differentiating intelligibility groups. Studies have demonstrated that consistent speech sound production/error(s) have a significant impact on intelligibility [111], and that some phonemes have significantly higher frequency of use and a larger role in phonemic contrasts and intelligibility (e.g. a high alveolar nasal and alveolar lateral load across languages with differing phoneme inventory types [112]). Therefore, the correct PPG probability for each instance of a phoneme aligned token ($\bar{\mathbf{g}}_p$) (cf. Section 2.2.2 for details) is plotted by manner of articulation in Figure 11 to enable analysis of PPG probability, articulation consistency, and speech intelligibility. Table 11 shows the results for pairwise KS tests for intelligibility groups by manner of articulation. The pairwise KS tests show significant differences in the distribution of correct probability values across all 18 pairwise comparisons ($\alpha < 0.05$). Statistically significant differences for all 6 manner of articulation groups were found for the comparison between Control vs. 'A' speakers (P≤ 0.0013), 'A' vs. 'C' speakers (P≤ 0.001), and 'C' vs. 'D/E' speakers (P≤ 0.028). Lateral approximants consistently showed the largest KS statistic (D) between Control vs. 'A', 'A' vs. 'C' and 'C' vs. 'D/E' groups ($D = 0.705, 0.554, 0.373$, respectively) indicating that the distribution of lateral approximant correct probability scores are most separated in the data. Affricates also show a large relative difference across all groups ($D = 0.253, 0.585, 0.277$, respectively). Approximants show a small relative difference for the 'C' vs. 'D/E' group, and a large relative difference between all other groups, and plosives show medium to medium-large relative differences. Finally, nasals consistently show the smallest relative difference ($D = 0.107, 0.136, 0.123$, respectively), indicating that the distribution of nasal correct probabilities is most similar between successive groups.

Table 11: Pairwise intelligibility group Kolmogorov-Smirnov test results by manner of articulation (correct probability values across all instances of phoneme aligned tokens ($\bar{\mathbf{g}}_p$)).

G1	G2	Manner	KS Statistic	KS P-value
Con.	'A'	plosive	0.218	< 0.001
Con.	'A'	nasal	0.107	< 0.001
Con.	'A'	fricative	0.120	< 0.001
Con.	'A'	affricate	0.253	0.0013
Con.	'A'	approximant	0.244	< 0.001
Con.	'A'	lat. approx.	0.705	< 0.001
'A'	'C'	plosive	0.116	< 0.001
'A'	'C'	nasal	0.136	< 0.001
'A'	'C'	fricative	0.416	< 0.001
'A'	'C'	affricate	0.585	< 0.001
'A'	'C'	approximant	0.319	< 0.001
'A'	'C'	lat. approx.	0.554	< 0.001
'C'	'D/E'	plosive	0.220	< 0.001
'C'	'D/E'	nasal	0.123	0.007
'C'	'D/E'	fricative	0.215	<0.001
'C'	'D/E'	affricate	0.277	0.028
'C'	'D/E'	approximant	0.146	< 0.001
'C'	'D/E'	lat. approx.	0.373	< 0.001

In summary, speakers with and without intelligibility impairment are easily distinguished by PPG phoneme token accuracy but there were no statistically significant differences between Control speakers and 'A' rated dysarthric speakers, and between 'C' and 'D/E' rated speakers (the most severe groups available in the data). Both PPG correct probability and articulation consistency are required to distinguish intelligibility severity in the *TORGO* data, where the distribution of all manner of ar-

ticulation groups are significantly different between all severity groups. Lateral approximants and affricates show consistently large relative differences in the distribution of correct probability values between all severity groups. Therefore, all FDA grades are easily separated although there are no 'A/B'-'B/C' and 'C/D-'D' speakers in the data to compare.

5. Conclusion

Current approaches to ADA focus on the classification of broad labels (generally intelligibility, or intelligibility as an indicator of speech severity) which remain limited in clinical utility beyond a broad estimator of speech intelligibility. There is also growing concern that the classifiers are learning other aspects of audio rather than dysarthric speech features. Furthermore, the minimally detectable change or clinically important difference and measurement error have not been considered [28]. It can be observed from the *TORGO* data and analysis in this paper that speakers with the same neurological diagnosis and sub-type of dysarthria with the same level of intelligibility can have different dysarthric profiles. Therefore, accurately classifying broad labels is also limited in clinical utility for differential diagnosis and management. Furthermore, there is further complexity if e.g. there are similar levels of intelligibility between speakers of different dysarthria sub-types or differing or co-morbid aetiology underlying intelligibility impairment. Additionally, early symptoms differ in MSDs [113, 114].

As the preliminary steps towards automatic FDA assessment (and clinically interpretable dysarthria assessment), the analysis has demonstrated potential for PPGs to capture a dysarthric profile. PPG information can distinguish control and dysarthric speakers and provide information on dysarthric speech production. Table 12 summarises the dysarthric speech processes in relation to FDA aspects. Overall, analysis of the control data shows distinct PPG phoneme boundaries, with acoustic similarity showing sensitivity to voicing and place within manner of articulation groups (with relatively low probability values observed). For the dysarthric data, phonological processes related to deviant nasality, plosive voicing contrast, and lingual articulation are captured by PPG probability. Analysis shows some representation of labial dysarthric processes, but this was not straightforward to interpret. Importantly, the analysis highlights auditory perceptual features relevant to FDA scoring in the *TORGO*, and the dysarthric processes that contribute to speech intelligibility impairment. This increases focus and understanding of impairment across the speech subsystems (e.g. related to the FDA aspects) and clinical utility for dysarthria diagnosis and management.

For the aim of establishing the extent that information on dysarthric speech processes are encoded in PPG features, gold standard EMA phoneme and alignment information was used to extract phoneme windows for the analysis conducted. Studies have investigated the impact of forced alignment errors on pronunciation assessment [115], and future work will focus on the utility of PPGs without gold standard alignment data, and as an interpretable feature for ADA. Phoneme alignment for the *TORGO* dataset is well studied [116, 117, 118]. Additionally, current work on relevant automatic systems and posterior-based confidence scores rely on alignment and sub-word annotation, for example manual alignment and syllable level annotation for VSA [33], and forced alignment and phone-level transcription for articulatory GoP [119], respectively.

Auditory perceptual features for FDA scoring in the *TORGO*, and analysis of the dysarthric speech processes that are encoded in PPG features have been documented in this work. This is an exploratory step to create a framework for interpretable features for ADA as a prerequisite to future work on the automatic prediction of clinically interpretable dysarthria assessment. Also, future work will focus on comparing dysarthria profiles for differential diagnosis, e.g. distinguishing speech characteristics associated with different neurological aetiology and dysarthria subtypes. Furthermore, as studies commonly use speech audio from the *TORGO* dataset, and due to the large scope of the paper, this work focused on audio with speech only and 'in speech' FDA subscore ratings. The FDA also includes non-speech tasks that contribute to dysarthria diagnosis. For example, performance on sequential and alternating diadochokinetics (DDKs) [120], and the profile of relative scores between speech and non-speech tasks [34]. Future work on non-speech tasks should use the appropriate FDA subscores and conduct analysis in context of the FDA aspects and dysarthria profile compiled.

The contextual effect on phonemes has been shown to impact on pronunciation models (including the spectral relationships) and as the *TORGO* data contains a limited pool of transcripts an assumption that factors of phoneme environment and contextual variability are similar across speakers has been taken in this work unless otherwise stated. The analysis highlights sensitivity to coarticulation and assimilation in PPG features for control and dysarthric speakers, indicating further exploration of classification features in dysarthric motor control and contextual variability may be beneficial. Also, the analysis highlights that although speakers F03 and F04 have 'A' rated intelligibility, there are differences in the lingual profile, and processes secondary to contextual variability (with potentially more distinction in some processes due to over-articulation [121]) relative to control speakers. However, further work is required to confirm these trends. Finally, as 8 of the 9 FDA aspects focus on abnormal characteristics of speech production (the remaining being intelligibility) the linguistic and semantic context was not considered for the scope of the paper (i.e. the characterisation of the dysarthric features (and therefore dysarthria profile) that underlie the changes in phoneme distinction). Notably, in the first edition of the FDA (which was administered in the *TORGO*), all stimuli in the sentence intelligibility subtest have a uniform structure consisting of an identical phrase subject and verb, with a variable subject complement. Additionally, the inclusion of phonemically contrastive complements and consistent morphological suffixes minimise contextual cues. An analysis of linguistic context and FDA intelligibility ratings could be beneficial in future work (although the intelligibility aspect stimuli were not recorded in the *TORGO* data which limits the scope of this analysis).

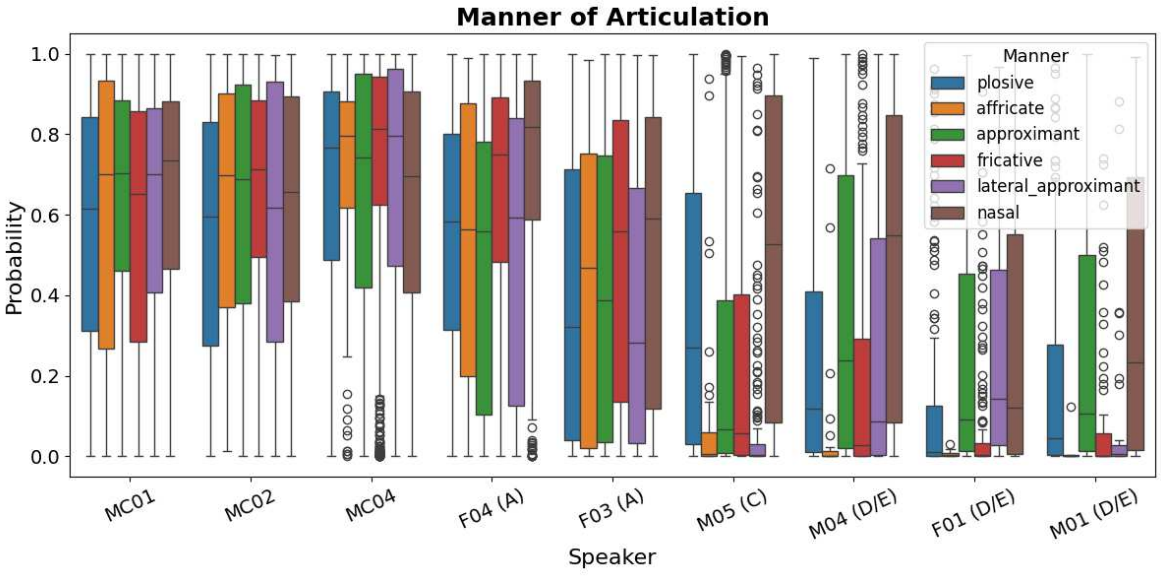

Figure 11: Correct Probability by Manner of Articulation for different speakers (sentence intelligibility ratings in parentheses).

Table 12: Dysarthric speech processes (also cf. Table 1) by PPG information.

FDA aspect	Summary
Respiration	Not examined. Information on e.g. syllables per breath unit and e/ingressive speech are not directly encoded in PPG tokens. Also, it is difficult to delineate distortion due to respiratory control with PPG information.
Jaw	Not examined (cf. Section 4.2 for details).
*Soft Palate	Speakers with and without FDA soft palate impairment are differentiated. Misclassification probability values are significantly different between severity groups for deviant nasality. Speaker with nasal emission differentiated (cf. Section 4.1).
*Tongue	Differentiates speakers with and without FDA tongue impairment. Distortion and substitution due to imprecise consonants: fronting/backing/stopping(/reduced frication), reduced retroflexion. Severity not examined (cf. Section 4.3).
*Lips	Distortion and substitution due to imprecise consonants: Labialisation(/approximant substitution)/reduced labiodental frication. Severity not examined. Not straightforward to interpret (cf. Section 4.4).
*Laryngeal	Voiced/voiceless plosive distinction. Differentiates speakers with and without FDA laryngeal impairment (cf. Section 4.5).
*Intel. (sentence)	Correct probability differentiates speakers with and without FDA intelligibility impairment. Statistically different correct probability values (across all aligned phoneme tokens $\bar{\mathbf{g}}_p$) between severity groups for all manner of articulation groups (cf. Section 4.6).

Acknowledgement

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] J. R. Duffy. *Motor Speech Disorders*. Elsevier, 2019.

[2] Petter Wannberg, Ellika Schalling, and Lena Hartelius. Perceptual assessment of dysarthria: comparison of a general and a detailed assessment protocol. *Logopedics Phoniatrics Vocology*, 41(4):159–167, 2016.

[3] Kate Bunton, Raymond D Kent, Joseph R Duffy, John C Rosenbek, and Jane F Kent. Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language and Hearing Research*, 2007.

[4] Abeer Muneer Altaher, Shin Ying Chu, Rogayah A Razak, et al. A report of assessment tools for individuals with dysarthria. *The Open Public Health Journal*, 12(1), 2019.

[5] Jessica Collis and Steven Bloch. Survey of uk speech and language therapists' assessment and treatment practices for people with progressive dysarthria. *International Journal of Language & Communication Disorders*, 47(6):725–737, 2012.

[6] Pamela Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[7] Nanako Hijikata, Michiyuki Kawakami, Ayako Wada, Maki Ikezawa, Kentaro Kaji, Yasuhiro Chiba, Miyuki Ito, Eri Fujino, Tomoyoshi Otsuka, and Meigen Liu. Assessment of dysarthria with frenchay dysarthria assessment (fda-2) in patients with duchenne muscular dystrophy. *Disability and Rehabilitation*, 44(8):1443–1450, 2022.

[8] Guilherme Schu, Parvaneh Janbakhshi, and Ina Kodrasi. On using the

ua-speech and torgo databases to validate automatic dysarthric speech classification approaches. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[9] Narendra Nonavinakere Prabhakera and Paavo Alku. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In *Interspeech*, pages 3403–3407. International Speech Communication Association (ISCA), 2018.

[10] Jinzi Qi et al. Speech disorder classification using extended factorized hierarchical variational auto-encoders. In *Proc. Interspeech 2021*, pages 1917–1921, 2021.

[11] F. Rudzicz, A. K. Namasivayam, and T. Wolff. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, Mar 2011.

[12] Subhashini Venugopalan, Jimmy Tobin, Samuel J Yang, Katie Seaver, Richard JN Cave, Pan-Pan Jiang, Neil Zeghidour, Rus Heywood, Jordan Green, and Michael P Brenner. Speech intelligibility classifiers from 550k disordered speech samples. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[13] Michael P Cannito, Debra M Suiter, Doriann Beverly, Lesya Chorna, Teresa Wolf, and Ronald M Pfeiffer. Sentence intelligibility before and after voice treatment in speakers with idiopathic parkinson's disease. *Journal of Voice*, 26(2):214–219, 2012.

[14] Kathryn M Yorkston, Edythe A Strand, and Mary RT Kennedy. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, 5(1):55–66, 1996.

[15] WHO. *International Classification of Functioning, Disability and Health: ICF*. Geneva :World Health Organization, 2001.

[16] Joan E Sussman and Kris Tjaden. Perceptual measures of speech from individuals with parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research*, 55(4):1208–1219, 2012.

[17] Katherine C Hustad. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *American Speech-Language-Hearing Association*, 2008.

[18] RCSLT. *Communicating Quality 3*. Oxon: RCSLT, 2006.

[19] Department of Health. *National Service Framework for Long-term Conditions (Neurological)*. HMSO, London, 2005. Product code: 26634.

[20] Anniek van Doornik, Marlies Welbie, Sharynne McLeod, Ellen Gerrits, and Hayo Terband. Speech and language therapists' insights into severity of speech sound disorders in children for developing the speech sound disorder severity construct. *International journal of language & communication disorders*, 60(3):e70022, 2025.

[21] Katharina Lehner, Wolfram Ziegler, and KommPaS Study Group. Indicators of communication limitation in dysarthria and their relation to auditory-perceptual speech symptoms: Construct validity of the kommpas web app. *Journal of Speech, Language, and Hearing Research*, 65(1):22–42, 2022.

[22] Kaila L Stipancic, Kira M Palmer, Hannah P Rowe, Yana Yunusova, James D Berry, and Jordan R Green. "you say severe, i say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research*, 64(12):4718–4735, 2021.

[23] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification using deep learning frameworks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 116–120. IEEE, 2021.

[24] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157, 2022.

[25] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, 158:103047, 2024.

[26] Shaik Sajiha, Kodali Radha, Dhulipalla Venkata Rao, Nammi Sneha, Suryanarayana Gunnam, and Durga Prasad Bavirisetti. Automatic dysarthria detection and severity level assessment using cwt-layered cnn model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):33, 2024.

[27] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu. Automatic assessment of dysarthria severity level using audio descriptors. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5070–5074. IEEE, 2017.

[28] Kaila L Stipancic, Yana Yunusova, James D Berry, and Jordan R Green. Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 61(11):2757–2771, 2018.

[29] Ghadeer Alharbi, Najwa Alamri, and Sahar Sabbeh. Automatic classification of speech dysarthric intelligibility levels using textual feature. *IEEE Access*, 2025.

[30] Macarious Hui, Jinda Zhang, and Aanchan Mohan. Enhancing aac software for dysarthric speakers in e-health settings: an evaluation using torgo. In *ICC 2025-IEEE International Conference on Communications*, pages 3673–3679. IEEE, 2025.

[31] Abner Hernandez, Sunhee Kim, and Minhwa Chung. Prosody-based measures for automatic severity assessment of dysarthric speech. *Applied Sciences*, 10(19):6999, 2020.

[32] Neethu Mariam Joy and Srinivasan Umesh. Improving acoustic models in torgo dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):637–645, 2018.

[33] Xiaokang Liu, Xiaoxia Du, Juan Liu, Rongfeng Su, Manwa Lawrence Ng, Yumei Zhang, Yudong Yang, Shaofeng Zhao, Lan Wang, and Nan Yan. Automatic assessment of dysarthria using audio-visual vowel graph attention network. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1454–1466, 2024.

[34] Pam Enderby and Rebecca Palmer. *Frenchay Dysarthria Assessment, Second Edition (FDA-2): Examiner's Manual*. Pro-ed, 2008.

[35] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[36] Amina Hamza, Djamel Addou, and Hamza Kheddar. Machine learning approaches for automated detection and classification of dysarthria severity. In *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM)*, volume 1, pages 1–6. IEEE, 2023.

[37] KL Kadi, SA Selouani, B Boudraa, and M Boudraa. Discriminative prosodic features to assess the dysarthria severity levels. In *Proceedings of the World Congress on Engineering*, volume 3, pages 1–5, 2013.

[38] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*, pages 1741–1744, 2008.

[39] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on advances in Signal Processing*, 2009:1–9, 2009.

[40] Dae-Lim Choi, Bong-Wan Kim, Yong-Ju Lee, Yongnam Um, and Minhwa Chung. Design and creation of dysarthric speech database for development of qolt software technology. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 47–50. IEEE, 2011.

[41] Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[42] Ray D Kent. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, 5(3):7–23, 1996.

[43] Abner Hernandez, Eun Jung Yeo, Sunhee Kim, and Minhwa Chung. Dysarthria detection and severity assessment using rhythm-based metrics. In *INTERSPEECH*, pages 2897–2901, 2020.

[44] Kamil Lahcene Kadi, Sid Ahmed Selouani, Bachir Boudraa, and Malika Boudraa. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering*, 36(1):233–247, 2016.

[45] HM Chandrashekar, Veena Karjigi, and N Sreedevi. Breathiness indices for classification of dysarthria based on type and speech intelligibility. In *2019 International Conference on Wireless Communications Signal*

Processing and Networking (WiSPNET), pages 266–270. IEEE, 2019.

[46] Ina Kodrasi and Hervé Bourlard. Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1210–1222, 2020.

[47] Parvaneh Janbakhshi and Ina Kodrasi. Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6477–6481. IEEE, 2022.

[48] Siddhant Gupta, Ankur T Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A Patil, and Rodrigo Capobianco Guido. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[49] Farhad Javanmardi, Saska Tirronen, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. Wav2vec-based detection and severity level classification of dysarthria from speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[50] Parvaneh Janbakhshi and Ina Kodrasi. Supervised speech representation learning for parkinson's disease classification. In *Speech Communication; 14th ITG Conference*, pages 1–5. VDE, 2021.

[51] Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, and Mateusz Lajszczak. Interpretable deep learning model for the detection and reconstruction of dysarthric speech. *INTERSPEECH 2019*, pages 3890–3894, 2019.

[52] Lingfeng Xu, Julie Liss, and Visar Berisha. Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1), 2023.

[53] Yishan Jiao, Visar Berisha, and Julie Liss. Interpretable phonological features for clinical applications. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 5045–5049. IEEE, 2017.

[54] James Carmichael. Diagnosis of dysayrthria subtype via spectral and waveform analysis. *Computer Systems Science & Engineering*, 29(1):33–42, 2014.

[55] Gary Weismer, Yana Yunusova, and Kate Bunton. Measures to evaluate the effects of dbs on speech production. *Journal of neurolinguistics*, 25(2):74–94, 2012.

[56] Brad H Story and Kate Bunton. Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):EL458–EL464, 2017.

[57] Robert Allen Fox and Ewa Jacewicz. Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in american english. *The Journal of the Acoustical Society of America*, 142(1):444–459, 2017.

[58] Jason A Whitfield and Daryush D Mehta. Examination of clear speech in parkinson disease using measures of working vowel space. *Journal of Speech, Language, and Hearing Research*, 62(7):2082–2098, 2019.

[59] Mousumi Malakar and Ravindra B Keskar. Progress of machine learning based automatic phoneme recognition and its prospect. *Speech Communication*, 135:37–53, 2021.

[60] Shobha Bhatt, Shweta Bansal, Ankit Kumar, Saroj Kumar Pandey, Manoj Kumar Ojha, Kamred Udham Singh, Sanjay Chakraborty, Teekam Singh, and Chetan Swarup. A comprehensive examination of phoneme recognition in automatic speech recognition systems. *Traitement du Signal*, 40(5), 2023.

[61] Wai-Kim Leung, Xunying Liu, and Helen Meng. Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136. IEEE, 2019.

[62] Xinwei Cao, Zijian Fan, Torbjørn Svendsen, and Giampiero Salvi. A framework for phoneme-level pronunciation assessment using ctc. *Interspeech*, 2024.

[63] Kavita Sheoran, Arpit Bajgoti, Rishik Gupta, Nishtha Jatana, Geetika Dhand, Charu Gupta, Pankaj Dadheech, Umar Yahya, and Nagender Aneja. Pronunciation scoring with goodness of pronunciation and dynamic time warping. *IEEE Access*, 11:15485–15495, 2023.

[64] Thomas Pellegrini, Lionel Fontan, Julie Mauclair, Jérôme Farinas, and Marina Robert. The goodness of pronunciation algorithm applied to disordered speech. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pages 1463–1467. ISCA, 2014.

[65] Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe, and David Mortensen. Leveraging allophony in self-supervised speech models for atypical pronunciation assessment. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:2613–2628, 2025.

[66] Silke M Witt and Steve J Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108, 2000.

[67] Timothy J Hazen, Wade Shen, and Christopher White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426. IEEE, 2009.

[68] Wei-Zhong Zheng, Ji-Yan Han, Chen-Kai Lee, Yu-Yi Lin, Shu-Han Chang, and Ying-Hui Lai. Phonetic posteriorgram-based voice conversion system to improve speech intelligibility of dysarthric patients. *Computer Methods and Programs in Biomedicine*, 215:106602, 2022.

[69] Wei-Zhong Zheng, Ji-Yan Han, Hsiu-Lien Cheng, Wei-Chung Chu, Ko-Chiang Chen, and Ying-Hui Lai. Comparing the performance of classic voice-driven assistive systems for dysarthric speech. *Biomedical Signal Processing and Control*, 81:104447, 2023.

[70] Wing-Zin Leung, Heidi Christensen, and Stefan Goetze. Text-to-dysarthric-speech generation for dysarthric automatic speech recognition: is purely synthetic data enough? In *International Conference on Speech and Computer*, pages 203–216. Springer, 2025.

[71] Gábor Gosztolya, Veronika Svindt, Judit Bóna, and Ildikó Hoffmann. Extracting phonetic posterior-based features for detecting multiple sclerosis from speech. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[72] Cameron Churchwell, Max Morrison, and Bryan Pardo. High-fidelity neural phonetic posteriorgrams. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 823–827. IEEE, 2024.

[73] Z. Yue, F. Xiong, H. Christensen, and J. Barker. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *ICASSP 2020*, pages 6094–6098, 2020.

[74] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In *Interspeech 2024*, pages 2494–2498, 2024.

[75] Anthony Seikel, Douglas King, and David Drumright. *Anatomy & Physiology for Speech, Language, and Hearing*. Thomson Delmar Learning, 2009.

[76] Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269, 1969.

[77] Ilias Papathanasiou and Patrick Coppens. *Aphasia and Related Neurogenic Communication Disorders*. JONES & BARTLETT PUB INC, June 2021.

[78] Joseph R Duffy and Raymond D Kent. Darley's contributions to the understanding, differential diagnosis, and scientific study of the dysarthrias. *Aphasiology*, 15(3):275–289, 2001.

[79] Yassin Khalifa, James L Coyle, and Ervin Sejdić. Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings. *Scientific Reports*, 10(1):8704, 2020.

[80] Kendrea L Garand, Gary McCullough, Michael Crary, Joan C Arvedson, and Pamela Dodrill. Assessment across the life span: The clinical swallow evaluation. *American Journal of Speech-Language Pathology*, 29(2S):919–933, 2020.

[81] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, 2019.

[82] John Kominek and Alan W Black. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, pages 223–224, 2004.

[83] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[84] Victor W Zue and Stephanie Seneff. Transcription and alignment of the timit database. In *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, pages 515–525. Elsevier, 1996.

[85] Jiatong Shi, Nan Huo, and Qin Jin. Context-aware goodness of pronunciation for computer-assisted pronunciation training. *INTERSPEECH2020*, pages 3057–3061, 2020.

[86] Nikola Maurová Paillereau. Do isolated vowels represent vowel targets in french? an acoustic study on coarticulation. In *SHS Web of Conferences*, volume 27, page 09003. EDP Sciences, 2016.

[87] Allard Jongman, Ratree Wayland, and Serena Wong. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263, 2000.

[88] Wiktor Gonet and Radosław Święciński. More on the voicing of english obstruents: voicing retention vs. voicing loss. *Research in Language*, 10(2):183–199, 2012.

[89] Wouter Jansen. Phonological 'voicing', phonetic voicing, and assimilation in english. *Language Sciences*, 29(2-3):270–293, 2007.

[90] Caroline L Smith. The devoicing of/z/in american english: Effects of local and prosodic context. *Journal of Phonetics*, 25(4):471–500, 1997.

[91] Kenneth L Moll and Raymond G Daniloff. Investigation of the timing of velar movements during speech. *The Journal of the Acoustical Society of America*, 50(2B):678–684, 1971.

[92] Fredericka Bell-Berti. Velopharyngeal function: A spatial–temporal model. In *Speech and language*, volume 4, pages 291–316. Elsevier, 1980.

[93] James Paul Dworkin, Mark T Marunick, and John H Krouse. Velopharyngeal dysfunction. *Perspective of the ASHA Special Interest Groups*, 2004.

[94] Lisa M Morris and Sherard A Tatum. *Craniofacial Surgery for the Facial Plastic Surgeon, An Issue of Facial Plastic Surgery Clinics*, volume 24. Elsevier Health Sciences, 2016.

[95] Antje S Mefferd, Abish Lai, and Francesca Bagnato. A first investigation of tongue, lip, and jaw movements in persons with dysarthria due to multiple sclerosis. *Multiple sclerosis and related disorders*, 27:188–194, 2019.

[96] Jimin Lee, Elizabeth Rodriguez, and Antje Mefferd. Direction-specific jaw dysfunction and its impact on tongue movement in individuals with dysarthria secondary to amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 63(2):499–508, 2020.

[97] Hermann Ackermann and Wolfram Ziegler. Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 54(12):1093–1098, 1991.

[98] Heejin Kim, Katie Martin, Mark Hasegawa-Johnson, and Adrienne Perlman. Frequency of consonant articulation errors in dysarthric speech. *Clinical linguistics & phonetics*, 24(10):759–770, 2010.

[99] Upashana Goswami, SR Nirmala, CM Vikram, Sishir Kalita, and SRM Prasanna. Analysis of articulation errors in dysarthric speech. *Journal of psycholinguistic research*, 49:163–174, 2020.

[100] RD Kent, HK Vorperian, JF Kent, and JR Duffy. Voice dysfunction in dysarthria: application of the multi-dimensional voice program™. *Journal of communication Disorders*, 36(4):281–306, 2003.

[101] Martin J. Ball. *Phonetics for Speech Pathology*. Communication Disorders and Clinical Linguistics Series. University of Toronto Press, third edition, 2020.

[102] Natalia Melle, Carlos Gallego, José María Lahoz-Bengoechea, and Silvia Nieva. Differential spectral characteristics of the spanish fricative/s/in the articulation of individuals with dysarthria and apraxia of speech. *Journal of Communication Disorders*, 109:106428, 2024.

[103] Sandra Clarke. *Newfoundland and Labrador English*. Edinburgh University Press, 2010.

[104] Ingo R Titze. *Workshop on acoustic voice analysis: Summary statement*. National Center for Voice and Speech, 1995.

[105] Shaheen N Awan and Nelson Roy. Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity. *JSLHR*, 2009.

[106] Rafael Monroy Casas and M Inmaculada Arboleda Guirao. Readings in english phonetics and phonology. *Institut Interuniversitari de les Llengues Modernes Aplicades (IULMA)*, 2013.

[107] Emily Fischer and Alexander M Goberman. Voice onset time in parkinson disease. *Journal of Communication Disorders*, 43(1):21–34, 2010.

[108] Hermann Ackermann and Ingo Hertrich. Voice onset time in ataxic dysarthria. *Brain and language*, 56(3):321–333, 1997.

[109] J Sean Allen, Joanne L Miller, and David DeSteno. Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 113(1):544–552, 2003.

[110] Gwen Van Nuffelen, Catherine Middag, Marc De Bodt, and Jean-Pierre Martens. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International journal of language & communication disorders*, 44(5):716–730, 2009.

[111] Mehmet Yavas and Regina Lamprecht. Processes and intelligibility in disordered phonology. *Clinical linguistics & phonetics*, 2(4):329–345, 1988.

[112] Caleb Everett. The similar rates of occurrence of consonants across the world's languages: A quantitative analysis of phonetically transcribed word lists. *Language Sciences*, 69:125–135, 2018.

[113] Arnold Elvin Aronson. Clinical voice disorders. *An interdisciplinary approach*, pages 157–197, 1985.

[114] Gabriela M Stegmann, Shira Hahn, Julie Liss, Jeremy Shefner, Seward Rutkove, Kerisa Shelton, Cayla Jessica Duncan, and Visar Berisha. Early detection and tracking of bulbar changes in als via frequent and remote speech analysis. *NPJ digital medicine*, 3(1):132, 2020.

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

[116] Yu Ting Yeung, Ka-Ho Wong, and Helen M Meng. Improving automatic forced alignment for dysarthric speech transcription. In *Interspeech*, pages 2991–2995, 2015.

[117] Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840. IEEE, 2019.

[118] Ying Li, Bryce Johannas Wohlan, Duc-Son Pham, Kit Yan Chan, Roslyn Ward, Neville Hennessey, and Tele Tan. Improving text-independent forced alignment to support speech-language pathologists with phonetic transcription. *Sensors*, 23(24):9650, 2023.

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

[120] Ray D Kent, Yunjung Kim, and Li-mei Chen. Oral and laryngeal diadochokinesis across the life span: A scoping review of methods, reference data, and clinical applications. *Journal of Speech, Language, and Hearing Research*, 65(2):574–623, 2022.

[121] Viviana Mendoza Ramos, Charlotte Paulyn, Leen Van den Steen, Maria E Hernandez-Diaz Huici, Marc De Bodt, and Gwen Van Nuffelen. Effect of boost articulation therapy (bart) on intelligibility in adults with dysarthria. *International Journal of Language & Communication Disorders*, 56(2):271–282, 2021.