PRONUNCIATION BASED INTELLIGIBILITY ASSESSMENT OF DYSARTHRIC SPEECH

Komal Bharti [*], *Ashita Batra* [†], *Pradip K. Das* [†], *Hemant A. Patil* [**]
komal.bharti@thapar.edu, (b.ashita, pkdas)@iitg.ac.in, hemant_patil@daiict.ac.in

[*]Thapar Institute of Engineering and Technology, Patila, India
†IIT Guwahati, Assam, India
**Dhirubhai Ambani University, Gandhinagar, Gujrat, India

ABSTRACT

The term dysarthria is composed of two parts: "dys," meaning *difficulty* and "arthr," refers to *articulation*. Individuals having dysarthria struggle a lot to control the movements of the speech-related muscles, resulting in reduced speech intelligibility, slow pace rate, lack of naturalness, and speech variability. With appropriate treatment, medication and consistent practice, the severity of speech difficulties can be alleviated. However, intelligibility assessment is a crucial and initial step that helps clinicians to understand the severity of the problem. We propose a speaker-independent method at the phoneme-level to identify the most frequently mispronounced phonemes among speakers in the UA-Speech database while keeping the computational complexity low. This method utilizes the Goodness of Pronunciation (GoP) algorithm, followed by GMM-GoP (Gaussian Mixture Model-GoP) and NN-GoP (Neural Network-GoP) for score calculations. We found that dysarthric subjects have difficulty in the articulation of phoneme /ŋ/, /æ/, /əʊ/, /ɪ/, and /aʊ/. Further, we calculated the correlation between the obtained GoP score and dysarthric speech intelligibility using Kendall's rank correlation coefficients and found a positive correlation 0.545 for GMM-GoP and 0.601 for NN-GoP, which establishes the reliability of the observed relationship.

Index Terms— Dysarthria, intelligibility assessment, phoneme boundary detection, GoP, phoneme score.

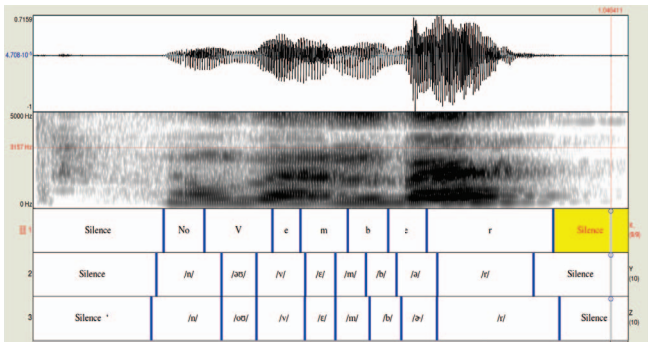
1. INTRODUCTION

Dysarthria is a motor speech disorder resulting from neurological damage to specific brain regions. Majorly caused due to severe accidents, strokes, or cerebral palsy. This damage disrupts the coordination between speech muscles and the tongue, resulting in a loss of control over articulation-related muscles. Consequently, individuals experience imprecise speech, a slow pace and unintelligible speech. Early checkups with appropriate treatment and regular exercises can significantly reduce the severity-level of dysarthria. In the traditional subjective method, speech intelligibility is assessed perceptually using articulation tests and manual acoustic analysis, which are costly, laborious, and subject to many listener biases. It also depends heavily on the previous experience, expectations, and perceptual skills of speech pathologist, which may lead to disparity among different speech pathologist. Due to the limitations in traditional evaluation methods, there is a need for reliable, automatic intelligibility assessment techniques for dysarthric speaker [1].

There are two primary methods for automatic dysarthric speech assessment. The first method looks into hand-crafted speech features [2], and the second method employs neural networks method focused on raw speech data for prediction. Dysarthric speech often exhibits distorted pronunciation at the phoneme-level, such as substitutions, omissions or distortions of specific sounds. By assessing each phoneme, the GoP algorithm identifies the phonemes that are mostly deteriorated with time. One distorted phoneme in a word is sufficient to decrease the intelligibility rate for dysarthric speakers. Hence, it is essential to study the behavior of patients at phoneme-level to know, where the exact problem lies instead of generalization of all the features. In this paper, we explore the GoP algorithm for dysarthric speech, which analyzes phonemes that are distorted and the extent each phoneme is atypical. This approach does not require large parallel datasets for training, making it well-suited for dysarthric speech assessment. Otherwise data scarcity and significant speaker variability remain major challenges in this domain [3]. Although GoP is frequently used to evaluate non-native speech pronunciation [4, 5], some research has confirmed that it may be useful in evaluating speech disorders too [6, 7, 8]. We utilized the GoP algorithm and calculated the final GoP score using both GMM-GoP and NN-GoP rather than the baseline GoP algorithm.

2. BACKGROUND

The earlier studies attempt to combine dysarthric speech assessment, and the utility of computer-based speech recognition in a single model [9]. These studies emphasize feature representation and prediction using a phonologically structured sparse linear model [10, 11, 12]. They often miss subtle patterns present in speech signals. In several studies, re-



(a)

(b)

Fig. 1. (a) TextGrid File with different tiers of annotation for the word "November", and (b) process of manual annotation for third tier.

searchers extract different types of feature vectors from the raw signal, such as prosodic features, acoustic and lexical features, vowel articulation features, audio descriptors and multitapered spectral estimation, and classify them using genetic algorithms or Artificial Neural Networks (ANN). A recent study [13] enhanced the GoP score using uncertainty quantification with the prior-normalized maxlogit GoP. The experiment was conducted across English, Korean, and Tamil languages. This study inspired us to conduct a detailed analysis of dysarthric speech at the phoneme-level.

For dysarthric speech, only three publicly available datasets are there: UA-Speech [14], TORGO [15] and Nemours. Due to the limited number of speakers in these datasets, researchers are often forced to focus on a speaker-dependent framework. In this paper all experiments are conducted using the UA-Speech database. Each speaker was assigned to one of four intelligibility categories based on perceptual evaluation, as presented in Table 1. Based on availability, we selected 13 healthy and 13 dysarthric speakers for experiments.

3. GOODNESS OF PRONUNCIATION (GOP)

The GoP algorithm estimates how closely a spoken phoneme matches its canonical form by computing a probabilistic likelihood ratio. The purpose of the GoP measure is to generate

Table 1. Database Description of UA-Speech Database

Categories	Speaker	Intelligibility %
Very Low	M01, M04, F03	(0-25)
Low	M06, M07, F02, M16	(25-50)
Mid	M05, M11, F04	(50-75)
High	M08, M09, F05	(75-100)

a score for each phoneme within an utterance [16]. We apply this algorithm to dysarthric speech, where each speaker exhibits unique characteristics. Therefore, we define the Goodness of Pronunciation (GoP) as the degree of similarity between the produced and the correct phoneme pronunciation in terms of the smoothness of articulation and how easily the pronunciation can be understood. For each phone q with its corresponding acoustic segment $O^{(q)}$, there is a set of HMMs available to estimate the likelihood that is $P(O^{(q)}|q)$. The quality of pronunciation of any phoneme p is defined as the log of the posterior probability given the corresponding acoustic segment $O^{(p)}$. In particular,

$$GoP(p) = \left|log(P(p|O^{(p)}))\right|/N_f(p), \quad (1)$$

$$GoP(p) = \left|log\Big[\frac{(P(O^{(p)}|p)P(p)}{\sum_{q\in Q} P(O^{(p)}|q)P(q))}\Big]\right|/N_f(p), \quad (2)$$

where $N_f(p)$ is the number of frames in the acoustic segment $O^{(p)}$. Q is set of all phonemes and $P(p)$ is the prior probability of phoneme p. It reflects how likely the phoneme p is to appear in general speech. The GoP algorithm consists of three key steps: (1) a forced phoneme alignment step, (2) a free phoneme recognition step, and (3) a scoring step, where score is calculated each forced aligned phone as the difference between the log-likelihoods of the two preceding steps. A higher absolute score signifies a larger difference between ideal and actual pronunciation. The architecture of the proposed work is shown in Fig. 2 where X_n is audio file, and T_n is its respective transcription. Experimental details of all the components of GoP is explained in the following sub-sections.

3.1. Forced Phone Alignment Phase

During this phase, the system is directed to match the spoken words with their expected sounds, ensuring that the timing and pronunciation align correctly. It uses a pronunciation dictionary to break words down into phones and aligns them with specific moments in the recording. Basically, it matches orthographic transcriptions to spoken audio at the phoneme-level. Initially, we had raw audio data with a list of words describing what is being spoken in those audio files. We select first 100 words from the UA-Speech dataset for all the speakers, which is sufficient to cover all the phonemes. The selection of words is same for all the speakers. For each audio file, its corresponding textGrid file is generated using Montreal Forced Aligner (MFA) with International Phonetic Alphabet (IPA) notation. Each TextGrid file contains temporal

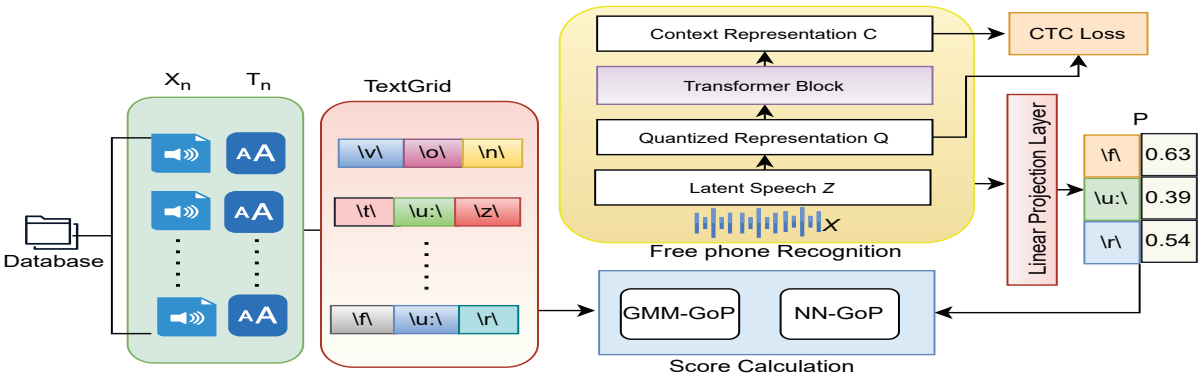


Fig. 2. Architecture of the proposed framework.

information delineating the duration covered by the phonemes having *start* and *end* times with "Min" and "Max" markers. These values corresponding to each phoneme represent the actual start and end time of occurrence of that event within the acoustic signal. During the forced alignment phase, the phoneme boundaries from the TextGrid are used to estimate the likelihood of acoustic features for each phoneme which is critical for accurate GoP computation. As shown in Fig. 1(a), the TextGrid file consists of three tiers each representing a distinct layer of annotation and the data from the third tier, annotated by an expert is used to extract the "Min" and "Max" values for further experimental settings.

The novelty of our work lies in the pre-processing part, primarily in the *manual* correction of phoneme boundaries for low and very-low intelligible speech. Automatic alignment using MFA is often inaccurate for low intelligible speech due to large pronunciation variability and acoustic mismatch. GoP scores are highly sensitive to the boundary errors. Therefore, phoneme boundaries were manually verified with the help of a speech expert as shown in Fig. 1(b). Word "November" is annotated in three stages. Tier 1 is by naive listener, tier 2 by MFA and tier 3 by a speech expert. In this way, we create a Textgrid file T for each audio file X.

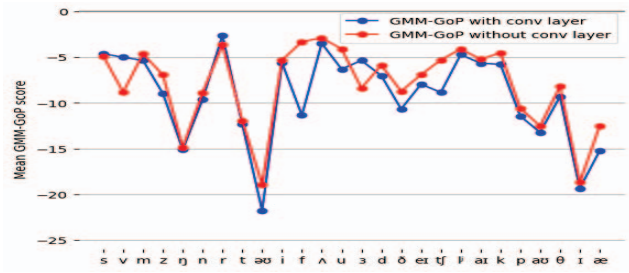
3.2. Free Phone Recognition Phase

This phase recognizes the phoneme sequence of the input speech signal without using any reference transcription and estimate log probability. It tries to determine the phoneme sequence that was actually spoken, purely based on the acoustic features learned from healthy speech. For data preparation, the TextGrid file is first converted to a structured CSV format that contains phone labels and the corresponding phoneme boundaries for each audio file [1]. For phoneme representation learning, Wav2Vec2.0 model is utilized. The model is trained to operate directly on raw waveform data instead of using traditional features, such as MFCC (Mel Frequency Cepstral Coefficients) or spectrograms, since we need raw pronunciation without any alteration. In this experiment, we selected the first 100 words (out of 455), ensuring comprehensive coverage of all the phonemes. Each word has 7 utterances available, so we used a total of $(7 * 100 * 13) * 2 = 18,200$ audio files for training. The model is trained on both healthy and dysarthric speech data. A stack of convolutional layers takes raw audio X and converts it to latent speech representation $Z_1, Z_2, ..., Z_T$ for T time steps. This process transforms the raw audio into a sequence of latent feature vectors, where each vector corresponds to a small segment of the original audio. They are then fed to the transformer to build representation $C_1, C_2, ..., C_T$ to capture contextual information and dependencies over the entire sequence. The output of the multi-layer convolutional feature encoder is discretized to q_t with a quantization module Q to represent the targets. In addition, linear projection layer is used for dimensionality reduction before passing it to the downstream classifiers. We fine-tuned the model according to dysarthric speech and its labeled data, followed by adding a classifier at the top of the Wav2Vec2.0 model. This layer maps the contextual representations to phoneme labels. We use the standard *train* and *test* split and follow the standard protocol of collapsing phone labels to 26 classes. The model is 82.3% accurate to predict phonemes accurately. The model is trained on both healthy and dysarthric speech data so that the model learns to extract useful features from the audio by predicting masked portions of the input.

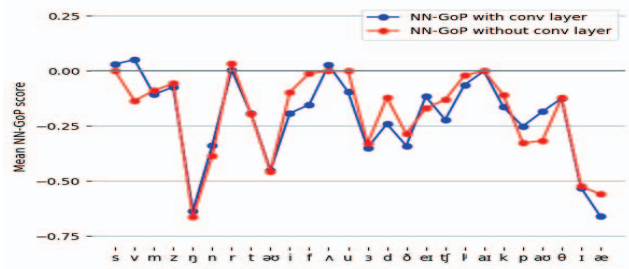
3.3. Score Computation

We measure intelligibility scores cluster-wise, for each phoneme using GMM-GoP and NN-GoP. Due to the significant variation in speaker characteristics, we divided them into two cluster based on intelligibility. We combined low and very low-intelligible speakers into one category, while the remaining speakers with mid and high intelligibility were placed in another one. GMM-GoP is effective for handling low-intelligibility speakers as it captures broad statistical patterns, whereas NN-GoP offers better flexibility and adapts to

[1] https://bit.ly/3P888Rv



(a)



(b)

Fig. 3. GoP score for each phoneme, where larger difference from baseline indicates potential mispronunciation. A higher absolute value reflects more deviation from the actual pronunciation. (a) GMM-GoP, and (b) NN-GoP.

variability in speech, especially for mid-to-high intelligibility speakers. Combining GMM-GoP and NN-GoP leverages the strengths of both approaches and helps to address the full spectrum of dysarthric speech.

Fig. 3 presents the mean GoP scores for each phoneme across all the speakers. As shown in Fig. 3(a), the majority of phoneme scores fall within the range of -5 to -20. In comparison, GMM-GoP exhibits a wider score range, reflecting higher severity and greater variation in pronunciation quality. The score range for NN-GoP varies with severity but is generally more concentrated than that of GMM-GoP. As shown in Fig. 3(b), NN-GoP produces both positive and negative scores, with most phoneme scores clustered near zero with narrow range, indicating relatively better pronunciation. In both cases, we identified a set of phonemes that dysarthric speakers find most challenging. For GMM-GoP, we found /əʊ/, /ɪ/ and /ŋ/ with higher score and for NN-GoP /ŋ/, /æ/ and /ɪ/ was most distorted. Table 2 provides details on the distorted phonemes along with their GoP scores. The GoP score is calculated by comparing the log-likelihoods of phoneme sequences obtained from the forced alignment phase and the free phone recognition phase.

4. RESULTS AND DISCUSSION

To make sure that the obtained phoneme score is associated well with dysarthric speech intelligibility, we compute the correlation between the obtained GoP score and dysarthric speech severity rate using Kendall's rank correlation coefficient (τ). As shown in Table 3, both GMM-GoP and NN-GoP systems show *positive* correlations between GoP scores and speaker intelligibility. A higher τ indicates that the proposed method effectively preserves the relative ranking of phonemes, demonstrating better alignment with speech intelligibility.

Table 2. Most distorted phonemes in ascending order associated with each cluster

Method	Speaker	Phoneme	Score
GMM-GoP	M01, M04, F03, M06, M07, F02, M16	əʊ	**-18.9067**
		ɪ	-18.6762
		ŋ	-14.8451
		aʊ	-12.5230
		æ	-12.5273
		t	-11.9644
		p	-10.5817
NN-GoP	M05, M11, F04, M08, M09, F05	ŋ	**-0.6610**
		æ	-0.6601
		ɪ	-0.5302
		əʊ	-0.4566
		n	-0.3850
		p	-0.3275
		əʊ	-0.3174

Table 3. Kendall's rank correlation coefficients for each intelligibility cluster

Method	Intelligibility	τ	Overall τ
τ **(GMM-GoP)**	Very Low	0.538	**0.545 (+ve)**
	Low	0.551	
τ **(NN-GoP)**	Mid	0.561	**0.601 (+ve)**
	High	0.640	

Table 2 reflects that the speakers with low intelligible speech tend to have lower GoP scores (indicating greater pronunciation difficulties) compared to mid and high intelligible speakers. We found that certain phonemes have more impact on the severity-level of the speaker. Despite several challenges in dysarthric speech, such as articulatory imprecision, abnormal prosody, phoneme confusion, and inherent acoustic differences, a positive correlation is found in our study for all intelligibility. When we repeated the same experiment on *healthy* speech for both training and testing, τ increased to **0.654**. This suggests that the train-test mismatch is significantly for dysarthric speech than for healthy speech. With this approach, we achieve Kindall's rank correlation coefficients close to the results observed for healthy speech, despite strong inter and intra-speaker variability.

In the previous study, the SOTA Kendall's (τ) for UA-Speech dataset is found to be 0.623 using MixGoP, and for helathy speech it is 0.539. They didn't provide intelligibility-wise Kendall's coefficients [17]. In comparison to the previ-

ous studies, our work *strongly correlated* with the intelligibility associated with each cluster and focuses on phoneme-level distortion analysis across clusters. Our contribution is not modifying GoP, however, analyzing its behavior and limitations when applied to dysarthric speech.

5. CONCLUSIONS

This paper presented the GoP algorithm for the assessment of dysarthric speech and identifies the most distorted phonemes in the UA-Speech database. Despite several challenges associated with dysarthric speakers, we successfully annotated all the audio files with accurate phoneme boundaries and developed a speaker-independent phoneme recognition system, followed by GoP score calculation. Positive Kendall's rank correlation coefficients reflect the strong association with the intelligibility score. We achieved a positive correlation while keeping the model simple and computationally efficient. This approach enables early identification of the problem at the phoneme level. However, the problem remains open for further refinement and optimization in future work.

6. ACKNOWLEDGMENT

The last author would like to acknowledge the partial support from MeitY, Govt. of India, under NLTM project 'BHASHINI', (Grant ID: 11(1)2022- HCC(TDIL)).

7. REFERENCES

[1] C. Bhat and H. Strik, "Speech technology for automatic recognition and assessment of dysarthric speech: An overview," *Journal of Speech, Language, and Hearing Research*, vol. 68, no. 2, pp. 547–577, 2025.

[2] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 02 2000.

[3] K. Bharti, S. Haque, and P. K. Das, "A novel dysarthric speech synthesis system using Tacotron2 for specific and OOV words," in *International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, 2024.

[4] J. Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," *The Journal of the Acoustical Society of America (JASA)*, vol. 134, pp. 1336–47, 08 2013.

[5] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *Proceedings of the ISCA Special Interest Group:SLaTE*, pp. 55–58, 2009.

[6] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, C. Alazard-Guiu, M. Robert, and P. Gatignol, "Automatic assessment of speech capability loss in disordered speech," *ACM Transactions on Accessible Computing*, vol. 6, pp. 1–14, 05 2015.

[7] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pp. 1463–1467, 2014.

[8] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29–43, 06 2019.

[9] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *ICASSP, Taiwan*, pp. 4605–4608, 2009.

[10] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[11] A. Hernandez, H.-Y. Lee, and M. Chung, "Acoustic analysis of fricatives in dysarthric speakers with cerebral palsy," *Phonetics and Speech Sciences*, vol. 11, pp. 23–29, 09 2019.

[12] R. Kent, G. Weismer, J. Kent, and J. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *The Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–99, 12 1989.

[13] E. Yeo, K. Choi, S. Kim, and M. Chung, "Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification," in *Interspeech*, pp. 166–170, 08 2023.

[14] H. Kim and H.-J. *et. al*, "Dysarthric speech database for universal access research.," in *Interspeech*, vol. 2008, pp. 1741–1744, 2008.

[15] F. Rudzicz, "Using articulatory likelihoods in the recognition of dysarthric speech," *Speech Communication*, vol. 54, pp. 430–444, 03 2012.

[16] K. t. Sheoran, "Pronunciation scoring with goodness of pronunciation and dynamic time warping," *IEEE Access*, vol. 11, pp. 15485–15495, 2023.

[17] K. Choi and Y. *et. al*, "Leveraging allophony in self-supervised speech models for atypical pronunciation assessment," in *Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2613–2628, 2025.