ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095366

STATISTICAL ANALYSIS OF SPEECH DISORDER SPECIFIC FEATURES TO CHARACTERISE DYSARTHRIA SEVERITY LEVEL

Amlu Anna Joshy[1,3], *P. N. Parameswaran*[1,3], *Siddharth R. Nair*[1,3], *Rajeev Rajan*[2,3]

[1]College of Engineering Trivandrum, Thiruvananthapuram,
[2]Government Engineering College, Barton Hill,
[3]APJ Abdul Kalam Technological University, India.

ABSTRACT

Poor coordination of the speech production subsystems due to any neurological injury or a neuro-degenerative disease leads to dysarthria, a neuro-motor speech disorder. Dysarthric speech impairments can be mapped to the deficits caused in phonation, articulation, prosody, and glottal functioning. With the aim of reducing the subjectivity in clinical evaluations, many automated systems are proposed in the literature to assess the dysarthria severity level using these features. This work aims to analyse the suitability of these features in determining the severity level. A detailed investigation is done to rank these features for their efficacy in modelling the pathological aspects of dysarthric speech, using the technique of paraconsistent feature engineering. The study used two dysarthric speech databases, UA-Speech and TORGO. It puts light into the fact that both the prosody and articulation features are best useful for dysarthria severity estimation, which was supported by the classification accuracies obtained on using different machine learning classifiers.

***Index Terms*—** dysarthria severity estimation, paraconsistent feature engineering, statistical analysis

1. INTRODUCTION

The speech disorder arising from poor coordination of the speech production subsystems is referred to as dysarthria. The speech impairments exhibited by dysarthric patients are defined in different dimensions such as articulation, phonation, prosody, nasality and intelligibility in literature. Imprecise articulations due to the retardation of lip, jaw and tongue movements, and irregular glottal closure patterns resulting in breathy voice are top among the most evident dysarthria symptoms [1], [2]. Phonation features can define the monotonicity and tempo perturbations exhibited by the dysarthric patients [3]. Dysarthric speech is often emotionless and lacks rhythm due to the abnormal speech rate and irregular loudness, and the prosodic measures can characterise them [3]. When associated with any degenerative disorders of the central nervous system and/or hereditary conditions, dysarthria can be progressive in nature. This demands the need for frequent monitoring of the severity level for proper medication and voice treatment during rehabilitation. However, subjective evaluation of the same by speech-language pathologists (SLP) would be biased, time-consuming, and expensive. Different approaches for automating this severity estimation are adopted in the literature. While the earlier works concentrated on feature selection [4], [5] and handcrafted feature generation [6], [7], more recent works focus on building end-to-end systems or sophisticated deep learning models with basic acoustic features [8], [9], [10], [11]. However, training deep learning models is prone to overfitting as the amount of dysarthric data available is limited. The physical fatigue and vocal strain faced by the dysarthrics lead to this challenge of data scarcity.

Our initial experiments using these speech disorder-specific features on deep neural networks(DNN) [12] suggested that a detailed statistical analysis is required to understand the potential correlation within each class. This would also enable a choice of the optimum feature descriptor that could be used by a simple predictor for aiding SLPs. When end-to-end systems aim to replace the need of an SLP, at the cost of data gathering requirements and computational costs, simple predictors such as a random forest(RF) classifier can aid SLPs after selecting an optimum descriptor. We implement the recently proposed technique of paraconsistent feature engineering(PFE) [13] to picture the intra-class similarities and the inter-class distinctions exhibited by these features. PFE is not exactly a statistical tool, but a similar data analysis tool that helps to draw meaningful conclusions from the features representing raw data. The descriptive nature of the statistical analysis is not shown by PFE as it does not uncover the structure behind the data. However, the exploratory nature is present implicitly as it helps in understanding the potential correlation among the features and the way they are mapped to the correct severity level. PFE has been shown to be efficient in feature ranking for applications such as replay attack detection [14] and speaker verification [15].

The proposed approach is explained in Section 2, followed by Section 3 describing the databases. The experimental framework and result analysis are given in Sections 4 and 5 respectively. Finally, the paper is concluded in Section 6.

2. SYSTEM DESCRIPTION

With the aim of analysing the potential of prosody, articulation, phonation and glottal-based features in recognising the paralinguistic aspects from the dysarthric utterances, we extract them and do the PFE analysis to rank their efficacy. Details of this are given below.

2.1. Disvoice feature set

Prosodic information is important in dysarthric characterisation as it can highlight the abnormalities in the intonation (pitch alterations) and voicing style (irregular phoneme and syllable durations). As explained in [1], 103 features based on duration, pitch and energy contour statistics are calculated. These include the linear estimation of the fundamental frequency(F0) and energy contour(cont.) over all segments, and calculation at the first(F0-F) and last(F0-L) voiced segments, duration analysis of voiced/unvoiced/pause segments, and their ratios. These can depict the level of monotonicity and the maximum frequency attainable by the patient [1].

Articulatory deficits are exhibited when there is a difficulty in changing the position/shape of the organs/tissues/limbs involved in the speech production [1]. The stress involved during the articulatory control and the resulting impairments can be understood by studying the onset(ON) and offset(OFF) transitions. Analysing the frequency content in these transitions can effectively model their difficulty in controlling the vocal fold vibrations [1]. Hence, Bark-band energies(BBE), the first two vocal formants(F1 and F2) and mel-frequency cepstral coefficients(MFCCs), along with their derivatives are extracted to account for 122 descriptors. To obtain a dynamic representation, the four statistical functionals, namely, mean(μ), standard deviation(σ), kurtosis(κ), and skewness(γ) are calculated on each feature per recording, thus giving the 488-dimensioned articulatory feature set.

Abnormalities in the phonation due to irregular glottal closure patterns give an impression of breathy voice to the speech [16] of people affected by hypokinetic dysarthria, as in the case of Parkinson's disease(PD). The glottal flow patterns are estimated using the glottal inverse filtering(GIF) technique called the iterative and/or adaptive inverse filtering(IAIF) as described in [2]. The time variability between consecutive glottal closure instants(GCI), the average and variability of features namely, the open quotient(OQ), the amplitude quotient(AQ), the normalized AQ with respect to glottal period(NAQ), the difference of the first two harmonics of the glottal flow signal(H1H2), and the harmonic richness factor(HRF) are calculated. Then the statistical measures are estimated, resulting in 36 features per utterance.

The phonatory measures could define the irregularities in the stability and periodicity of the vocal fold vibrations shown by the dysarthrics [17]. The long–term variabilities in the peak-to-peak amplitude and pitch(perceived F0) are measured in terms of perturbation measures such as jitter, shimmer, amplitude perturbation quotient(APQ), and pitch perturbation quotient(PPQ). The first and second derivatives of F0, and the logarithmic energy are also calculated to understand the spread of F0, which would indicate the measure to which the tongue and velum can be controlled [1]. The statistical functionals when applied to these seven features give the 28 dimensional phonation feature set.

2.2. Paraconsistent feature engineering(PFE)

To quantitatively compare the utility of the four feature sets, we adopt the PFE framework [13] for statistical analysis. As the first step, whole of the available X number of feature vectors are L2-normalised to the range [0, 1]. Then the intra-class similarities and the inter-class dissimilarities are analysed using the parameters, α, the level of faith and β, the level of discredit respectively. Consider a 1-D feature representation with the difference between the maximum possible value and the minimum value within a specific class being A. Advantageous intra-class similarity occurs when A is small, or $Y = 1 - A$ is large for each class. Thus, $Y = 0$ indicates low intra-class similarity and $Y = 1$ means high similarity. When the feature vector dimension is D and it is a N-class problem, the intra-class similarity can be quantified as the mean of Y calculated over each dimension. Then, the parameter α accommodates the worst-case scenario by selecting the minimum. These are computed as [13],

$$\bar{Y}_N = \frac{1}{D}\sum_{i=1}^{D} y_N(i) \tag{1}$$

$$\alpha = min\left\{\bar{Y}_1, \bar{Y}_2, ..., \bar{Y}_N\right\} \tag{2}$$

Next, the inter-class distinction is quantified using two range vectors for each class, containing the minimum and maximum values shown in each dimension over the entire dataset. Thus it gives the possible range of values within which the feature value is expected to lie for a given class. Now, the number of overlaps, Z is calculated as the count of features in one class lying within the range vector of all the other classes. This overlap is to be minimised and hence β is computed to be,

$$\beta = \frac{Z}{F} \tag{3}$$

where, F represents the maximum possible number of overlaps, which is $N.(N-1).X.D$ and $0 \leq \alpha, \beta \leq 1$. The final measures of the degree of certainty $G1$, and the degree of contradiction $G2$ are calculated respectively as,

$$G1 = \alpha - \beta,\ G2 = \alpha + \beta - 1; \quad -1 \leq G1,\ G2 \leq 1 \tag{4}$$

They define a new two-dimensional paraconsistent plane [13], where the ideal case of linearly separable features lies at the the corner (1, 0). Hence, the distance D between the point $P = (G1, G2)$ for each feature set to the ideal point (1, 0) can be calculated using the Euclidean distance to quantify its suitability for the chosen classification problem.

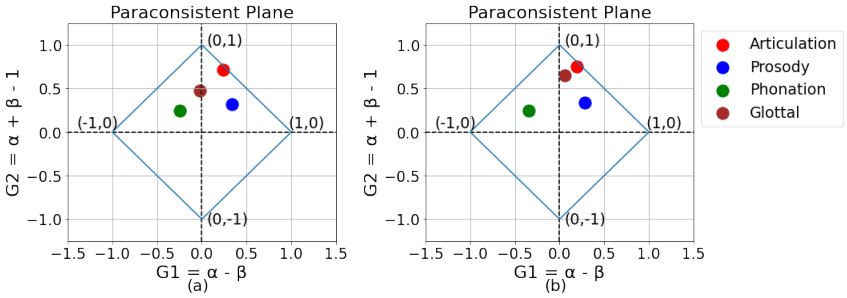

Fig. 1: Plots of different feature points in the paraconsistent plane (a) UA-Speech (b) TORGO

3. DATABASES

The English dysarthric speech databases, namely UA-Speech [18] and TORGO [19] are used for the analysis. Dysarthric speech of 15 patients from the former and eight patients from the latter is used. For training the machine learning(ML) classifiers, UA-Speech utterances corresponding to the 10 digits, 19 computer commands, 26 radio alphabets and 100 common words, all repeated thrice, are used. This sums to 465 recordings per speaker. The testing data has 300 distinct uncommon words per speaker. The severity of the disease `(HIGH/MEDIUM/LOW/VERY LOW)` is determined based on the intelligibility assessment by five listeners. For the evaluation of TORGO database, the short word utterances are used. There are 2227 such utterances, and a 80%-20% train-test split is adopted. The severity of the speakers was assessed according to the standardized Frenchay dysarthria assessment. Detailed description is given in Table 1.

Table 1: Class-wise patient description

Severity	UA-Speech	TORGO
VERY LOW	F05, M08, M09, M10, M14	F03, F04, M03
LOW	F04, M05, M11	F01,M05
MEDIUM	F02, M07, M16	M01, M02, M04
HIGH	F03, M01, M04, M12	-

4. EXPERIMENTAL FRAMEWORK

The DisVoice [1] library and the KALDI toolkit are used for computing the features. Only the utterance level features are computed, and the PRAAT algorithm was employed for F0 calculation. Experimental analysis is done using PFE and the ML classifiers, namely support vector machine(SVM), Naive Bayes(NB), k-nearest neighbour(kNN) and RF classifiers. Tuning of the SVM model was done initially for linear and rbf kernels. Best among them was found to be linear, and hence further tuning was done with respect to the regularisation parameter c=1 to 10. RF and kNN were tuned for number of trees=10 to 200, and number of neighbours=10 to 100 respectively. Since, our aim is to do an analysis on the feature side and not to find the best classifier setting, we have not used a separate validation set during this tuning. Hence, values reported correspond to the best testing accuracy obtained.

[1] `https://github.com/jcvasquezc/DisVoice`

5. RESULTS AND ANALYSIS

The results obtained and inferences learned from the paraconsistent analysis of the Disvoice feature sets are briefed below.

5.1. PFE Analysis

PFE analysis of the different features on whole of the UA-Speech and TORGO databases resulted in the plots shown in Fig. 1(a) and Fig. 1(b) respectively. It can be found that the prosody feature set has the critical point P lying closest to the ideal point (1,0), followed in Euclidean distance by the articulation feature set for both the databases. The α, β and the distance D obtained for each of the feature set is tabulated in Table 2. We find that for both the databases, the articulation feature set gives the highest α value indicating that these feature vectors have high similarity within each severity level. But, on comparing the β and D values, the prosody ranks first indicating that their inter-class dissimilarity is the greatest and hence, can discriminate the dysarthria severity levels well. This suggests the usage of any simple easy-to-perform classifiers to be used on them for severity estimation. Whereas, if the phonation features are to be used, a strong/advanced classifier has to be used to effectively mitigate the misclassifications due to overlap of the inter-class feature vectors and scattering of the intra-class feature vectors.

Table 2: Paraconsistent framework on features (best values in bold)

-	UA-Speech			TORGO		
Feature set	α	β	D	α	β	D
Prosody	0.83	**0.49**	**0.73**	0.81	**0.53**	**0.79**
Articulation	**0.97**	0.74	1.05	**0.97**	0.78	1.11
Glottal	0.73	0.75	1.12	0.86	0.79	1.14
Phonation	0.49	0.75	1.27	0.45	0.79	1.36

5.2. Classifier analysis

In adherence to our findings reported in [12], the articulation feature set performed the best on most of the classifiers, as seen in Table 3. The efficacy of the articulation features over the rest has also been proved in the diagnosis of PD patients in [3] and [20]. However, it was found from the PFE analysis that they have lower inter-class dissimilarity than prosody. Dysarthria under progressive cases shows varying severity with time, but with varying patterns exhibited by speakers based on their underlying neurological cause. Thus

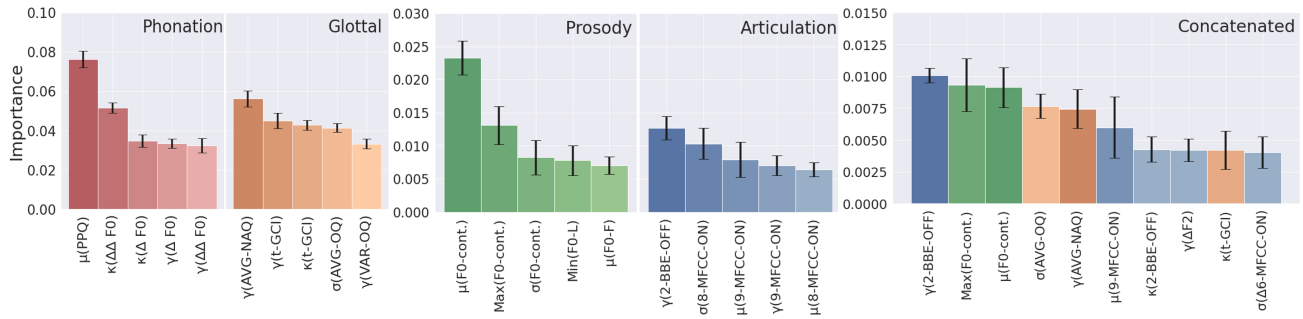

Fig. 2: Feature importance graph using permutation on the UA-Speech database(X-axis shows important features from each set)

the dysarthric speech is highly complex. Hence, the high intra-class similarity would have reduced the misclassifications considerably when the ML classifiers were used with the articulatory features.

When PFE claimed that prosody features require a relatively low-complex classifier in Table 2, it is found from Table 3 that, on TORGO database, these features work poorly on NB classifier compared to the rest. But the best accuracy given by the NB classifier using phonation features is just 54.29%, which does not guarantee an acceptable performance for real-time implementation. On the UA-Speech database, which is larger in terms of the number of speakers and total audio duration, articulation leads the rest, followed by prosody, but again with 54.02% accuracy only. This is due to the high variability of the dysarthric speech that cannot be modeled by the simple NB classifier. The zero frequency problem and the assumption of independence in the NB implementation led to its poor performance on all the features.

Table 3: Classification accuracy (%) obtained on different classifiers (best values in bold)

Database	Classifier	Phonation	Glottal	Prosody	Articulation
TORGO	SVM	62.88	55.60	60.18	**83.18**
	RF	69.14	76.45	81.49	**85.65**
	kNN	60.09	50.44	69.23	**73.99**
	NB	**54.29**	44.84	39.90	45.74
UA-Speech	SVM	60.81	55.91	61.68	**77.98**
	RF	65.82	70.86	67.72	**77.64**
	kNN	53.38	43.33	54.90	**60.69**
	NB	46.12	43.40	46.89	**54.02**

We calculated the feature importance within each feature set on the UA-Speech database (since it is the largest and has all four severity levels). This was done by noticing the increase or decrease in error when we permute the values of a feature. This approach is model-agnostic and does not have bias towards high cardinal features. The top 5 features in each set are depicted in Fig. 2. The graph also shows the standard deviation as error bars, whose length reveals the level of uncertainty. Since we have obtained short bars, the values are concentrated. The last plot gives the top 10 most important features from the concatenated feature set (obtained on concatenating the four feature sets) with a dimension of 655. It can be found that the most important features are from the articulatory, followed by prosody and glottal feature sets. The most discriminating feature is found to be the "skewness of 2^{nd} BBE on offset transitions", in the concatenated, as well in the articulation set. It has been shown in [1] that the BBEs are considerably reduced in dysarthics compared to healthy speakers. Now by PFE, we prove them to be best useful in differentiating the dysarthria severity levels as well. We find that the top four features in the best performing articulatory feature set are the MFCCs, which affirms the results of [10]. The change in Y axis values is due to the change in the fraction of contribution of each feature with increasing dimensionality. As reported in our earlier experiments [12] and supported by the findings in [21], the classification accuracy does not improve with the mere increment in feature dimension.

5.3. Discussion

To the best of our knowledge, this study is the first of its kind, in analysing the different speech disorder specific acoustic features for dysarthria severity classification using the PFE framework. Results report the usefulness of prosody and articulation features over the rest, and are supported by the classification accuracies obtained on using different ML classifiers. In the present era of deep learning, this study is relevant due to two reasons: (1) Taking into consideration the availability of low resource of impaired speech data, this analysis proved to be useful in demonstrating the efficacy of the different available features under data stringent conditions. (2) The analysis can be extended to other speech disorders like apraxia, and to specific cases of dysarthria like hypokinetic dysarthria exhibited by PD. The ranking of the features would be helpful in implementing simple predictors without the problem of over-fitting, to aid SLPs. Further, statistical hypothesis testing can be done on the classifiers, to find the optimum setup for dysarthria severity classification.

6. CONCLUSION

This paper presented a detailed analysis using paraconsistent framework to draw inferences about the performance of speech disorder specific features when used for classifying dysarthria severity levels. The results put light into the fact that a proper choice of features at the front-end by the PFE framework would enable the usage of simple predictors under the data stringent conditions.

7. REFERENCES

[1] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Process.*, vol. 77, pp. 207–221, 2018.

[2] Belalcázar-Bolanos *et al.*, "Glottal flow patterns analyses for parkinson's disease detection: acoustic and nonlinear approaches," in *Int. Conf. Text, Speech, and Dialogue*. Springer, 2016, pp. 400–407.

[3] J. C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *J. Commun. Disorders*, vol. 76, pp. 21–36, 2018.

[4] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Discriminative prosodic features to assess the dysarthria severity levels," in *Proc. World Congress on Engg.*, vol. 3, 2013.

[5] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *Proc. IEEE Int. Conf. Telecommun. Signal Process.*, 2016, pp. 515–518.

[6] H. Chandrashekar, V. Karjigi, and N. Sreedevi, "Breathiness indices for classification of dysarthria based on type and speech intelligibility," *Proc. IEEE Int. Conf. Wireless Commun. Signal Process. Network.*, pp. 266–270, 2019.

[7] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 3403–3407, 2019.

[8] H. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE J. Selected Topics Signal Process.*, vol. 14, no. 2, pp. 390–399, 2019.

[9] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using blstm," *IEEE J. Selected Topics Signal Process.*, vol. 14, no. 2, pp. 322–330, 2020.

[10] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 116–120.

[11] H. Tong, H. Sharifzadeh, and I. McLoughlin, "Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning," *Proc. Interspeech*, pp. 4786–4790, 2020.

[12] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1147–1157, 2022.

[13] R. C. Guido, "Paraconsistent feature engineering [lecture notes]," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 154–158, 2018.

[14] A. T. Patil, R. Acharya, H. A. Patil, and R. C. Guido, "Improving the potential of enhanced teager energy cepstral coefficients (etecc) for replay attack detection," *Computer Speech & Language*, vol. 72, p. 101281, 2022.

[15] A. M. G. de Almeida, C. H. Recco, and R. C. Guido, "Use of paraconsistent feature engineering to support the long term feature choice for speaker verification," *The International FLAIRS Conference Proceedings*, vol. 34, Apr. 2021.

[16] I. Midi, M. Dogan, M. Koseoglu, G. Can, M. Sehitoglu, and D. Gunal, "Voice abnormalities and their relation with motor dysfunction in parkinson's disease," *Acta Neurologica Scandinavica*, vol. 117, no. 1, pp. 26–34, 2008.

[17] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech," *Cognitive Computation*, vol. 9, no. 6, pp. 731–748, 2017.

[18] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *Ninth Annual Conf. Int. Speech Commun. Asso.*, pp. 1741–1744, 2008.

[19] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resources and Evaluation*, vol. 6, no. 4, pp. 523–541, 2012.

[20] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in parkinson's disease," *J. Voice*, vol. 25, no. 4, pp. 467–472, 2011.

[21] H. Holmström and V. Zars, "Effect of feature extraction when classifying emotions in speech-an applied study," *Ph.D dessertation, Umeà University, Sweden*, 2018.