DOI: 10.17469/O2111AISV000006

BARBARA GILI FIVELA, RAFFAELLO CASERTA, SONIA D'APOLITO,
ANNA CHIARA PAGLIARO, VINCENZO SALLUSTIO, MARCO SINISCALCHI

Phonetic analysis and deep neural network examination for dysarthria detection beyond sociolinguistic differences: a comparison

The goal of the paper is to compare a phonetic and an automatic evaluation, via Convolutional Neural Networks, of both dysarthric and control speech. On the one hand, we want to observe if and how phonetic and automatic classifications of dysarthric and healthy speech perform differently, given their specificities; on the other hand, we aim at checking if the automatic analysis is deceived by sociophonetic traits, taking them as instances of non-precise speech production. For the purpose of the analysis, we used a corpus of acoustic recordings of speech produced by 15 male parkinsonian dysarthric speakers (8 from Bari; 7 from Lecce) and 10 male control speakers from the same areas. Part of the corpus was used for fine-tuning the Neural Network that was previously trained on Spanish data; the remaining part of the corpus was used for testing the automatic and the phonetic evaluation. The comparison shows that the automatic and phonetic analyses seem to offer similar results in terms of classification, also with respect to sociolinguistic features. However, they could be successfully integrated as they suggest different information on the speakers and their speech.

Keywords: Tranfer learning, Italian, Spanish, individual strategies, sociophonetics.

1. *Introduction*

Speech accuracy is considered a good metric for assessing the severity level of motor speech disorders, such as dysarthria. Among the objective methods available to assess accuracy, phonetic analysis provides very detailed information, shedding light on individual strategies, even with reference to specific speech sounds or gestures. For instance, acoustic studies have shown that vowels produced by parkinsonian dysarthric speakers (PDs) occupy a reduced acoustic space compared to those of healthy controls (HCs), and show specific acoustic characteristics, such as increases in jitter and noise-to-harmonics ratio (a.o., Bang, Min, Sohn & Cho, 2013; Tjaden, 2000); as for consonants, they may be less precisely produced (a.o. Kim, Gurevich, 2021). Phonetic investigation results may be correlated to severity levels that are actually guaranteed by overall, usually subjective evaluations of dysarthria (e.g., Robertson's Dysarthria Profile; Robertson, 1982; Fussi, Cantagallo, 1999) or that may offer an overall and not speech-specific or dysarthria-specific evaluation (e.g., Unified Parkinson's Disease Rating Scale – UPDRS – and Hoehn and Yahr; Skodda, Gronheit & Schlegell, 2012; Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 2003). Nevertheless, in the case of phonetic

investigations, especially in the case of articulatory studies, analysed speech sounds or even populations may often not be wide enough to allow for robust generalization (Wong, Murdoch & Whelan, 2010; Wong, Murdoch & Whelan, 2011).

On the other hand, automatic learning algorithms are also used for objectively evaluating speech accuracy by considering global speech characteristics, even though no detailed observation on specific sound properties is provided (Wang, Fu, Shao, Chang, Ren, Chen & Ling, 2022, but see also Audibert, Fougeron, Fredouille, Meunier & Panseri, 2010). From a technical standpoint, the speech impairments exhibited by individuals with Parkinson's Disease (PD) have been analyzed by computing features associated with four dimensions: phonation, articulation, prosody, and intelligibility. Despite the established effectiveness of traditional feature extraction methods, recent years have witnessed the successful application of deep learning techniques for evaluating particular aspects of speech, such as the detection and monitoring of PD. These diverse speech analysis techniques encompass the utilization of technologies ranging from Support Vector Machines (SVM) to Convolutional Neural Networks (CNN). According to Vásquez Correa, Arias, Orozco-Arroyave & Nöth (2018), it is possible to achieve automatic recognition of speech characteristics by employing a CNN to analyze features extracted directly from the speech audio and corresponding to distinctive articulatory segments. This approach allows for the automated assessment of dysarthria severity through the examination of specific articulatory patterns realizing speech segments. However, it is noteworthy that the exploration of entire words or phrases remains an unexplored area in this context. Further research is required to delve into the potential insights that a comprehensive analysis of phonation and articulation across complete linguistic units could provide in understanding and assessing PD-related speech impairments.

A further issue regards sociophonetic features, which characterize everyone's speech since they are the manifestation of each speaker sociolinguistic identity, being due to various factors including the speech characteristics of the area from which the speaker comes from. Some sociolinguistic features may be the result of reduction or lenition phenomena that have taken place during the evolution of a specific dialect and may have an impact on the variety of the language spoken in the same area. Crucially, phenomena such as the reduction of vowels and the lenition of consonants may also be considered as the outcome of a diminished accuracy and precision in speech articulation in pathological speakers. In phonetic or subjective evaluation, they may be recognized as such, thanks to a phonology-driven approach to the phonetic analysis or to a sort of normalization made by speech therapists; but how do automatic analyses work in these cases?

Questions then arise concerning 1) the possible differences in efficacy and accuracy in distinguishing healthy and pathological speech by analyzing specific speech events, as in most of phonetic studies, or by exploiting automatic evaluations which consider the acoustic signal as a whole. In other words, one may wonder if the in-depth phonetic analysis of specific phenomena is representative of the overall speech accuracy or it rather describes too local events that do not necessarily fit

the severity level assigned to the experimental subject or to an entire sentence or item s/he produces; on the other hand, one may also wonder if 2) automatic analyses of speech, blind to very specific linguistic knowledge, may be deceived by sociophonetic traits, taking them as instances of non-precise speech production.

In this paper, we try to answer these two questions by comparing a phonetic (acoustic) analysis and an automatic evaluation of both dysarthric and control speech. This work is part of a larger PRIN project aiming at exploiting phonetic analysis of dysarthric speech by speakers of different varieties of Italian to develop clinical tools that face sociophonetic variation. Phonetic analyses are carried out semiautomatically (via scripts) and statistically, while automatic evaluations are carried out by means of CNNs (plus a check by means of SVMs), implemented on purpose for the project mentioned above.

The article is structured as follows: the state-of-the-art for investigations on dysarthric and control speech, with specific attention to the varieties considered here, is offered in § 2, and that for investigations based on deep neural networks for dysarthria detection is given in § 3; § 4 reports goals and hypotheses; the methodology is described in § 5, and the results of the present investigation are reported in § 6. The discussion and conclusions found in § 7 close the paper.

2. *Dysarthric and control speech in Lecce and Bari Italian*

As mentioned in § 1, in speech produced by PDs affected by (hypokinetic) dysarthria consonants are articulated with reduced accuracy and vowels are reported to occupy a reduced acoustic space compared to those of HCs, so that the vowel space could be a possible biomarker for the early detection of dysarthria (Roland, Huet, Harmegnies, Piccaluga, Verhaegen & Delvaux, 2023).

As far as Italian is concerned, experiments on acoustic and articulatory vowel space have shown that reduction takes place, but it is not homogenous (Gili Fivela, d'Apolito & Sigona 2022a). In particular, asymmetries in reducing the acoustic vowel space may be found, in line with articulatory data showing that some subjects appear to compensate in order to get a specific acoustic output. Concerning consonants, in line with results on other languages, the closure phase in plosives has been found to be not properly realized or, in any case, not articulated as HC do. However, plosives as well as fricatives are produced as accurately as possible as they are relevant phonemes in the languages. On the other hand, a sociolinguistic marker such as aspiration may be not preserved through compensatory strategies as much as the difference between phonologically relevant segments (Gili Fivela, d'Apolito & Pagliaro, 2022b). However, the disease seems to have a stronger impact on the production of the plosives than that of fricatives (Gili Fivela, Pagliaro & d'Apolito, 2023), even though correlates such as duration may change in both plosives and fricatives compared to HC segments (Gili Fivela et al., 2022b).

Crucially, varieties of Italian show different sociophonetic features that may be interestingly consider in the investigation of dysarthria. For instance, the varieties

of Italian spoken in Bari and Lecce (Southern Italy), belong to different dialectal isoglosses (South vs. Extreme South; Rohlfs, 1949) and are characterized by different sociophonetic markers, such as unvoiced plosive strong aspiration in Lecce, and vowel reduction or deletion in pre- or post-tonic position in Bari (where the system is eptavocalic, while it is pentavocalic in Lecce; Loporcaro, 1997; Romano, Manco, 2004). Thus, some of the characteristics variably distributed in the two varieties go in the direction expected in dysarthric speech, that are a variable phasing of laryngeal and supralaryngeal gestures (Best, 1995; Browman, Goldstein 1986), such as in unvoiced plosive aspiration in the Lecce variety, and a tendency to reduce vowels, such as in vowel reduction in unstress position in Bari Italian. Some research on the impact of sociolinguistic features on the dysarthria progression has been mentioned above and the issue is considered also in the analysis presented in the next sections.

3. *Deep neural networks for dysarthria detection*

The ultimate goal in using deep neural networks is the development of a diagnostic tool or diagnostic support for clinical personnel in the early diagnosis of PD through the automatic analysis of patients' speech and the identification of dysarthria symptoms. The system can provide an estimate of severity at the network's output, or it can simply binary classify characteristic features extracted from a suitable dataset. The desired diagnostic tool focuses precisely on analysing a whole recording, that is, a set of selected word or sentences. The objective is to prompt a patient to undergo an analysis of her/his speech under conditions compatible with a hospital setting, rather than in a laboratory specifically designed for the analysis of audio recordings.

Convolutional Neural Networks have proven to be among the most effective models for impaired speech analysis, as demonstrated by Vásquez Correa et al. (2018). CNNs belong to a class of deep neural networks designed to process grid-like data structures, such as images and, in the context of speech, spectrograms generated from audio signals. Their architecture features convolutional layers capable of automatically learning hierarchical representations, effectively capturing both local and global patterns. This capability makes CNNs especially well-suited for speech analysis tasks, where preserving the spatial relationships within input data is essential.

Despite their strengths, CNNs present certain challenges in speech analysis applications. Chief among these is the requirement for large volumes of labeled data for effective training, which can be both time-consuming and resource-intensive to obtain. Furthermore, transferability across datasets poses a concern, as CNNs trained on one dataset may not perform well on another related task. To address this, transfer learning — where a pre-trained model is fine-tuned on a new but related task — can be employed to leverage existing knowledge and improve generalization to new speech analysis domains.

4. *Aims and hypotheses*

The main goal is to observe if and how phonetic and automatic classifications of dysarthric and healthy speech perform differently, given their obvious specificities. A second goal is to check if automatic analyses of speech may be deceived by sociophonetic traits, taking them as instances of non-precise speech production.

The main hypothesis is that the phonetic and automatic procedures may classify speech consistently but do not allow for easy comparison, given the specificities mentioned in § 1. Still, they may be integrated, offering equally valuable, though somehow different, evaluations of the same speech. Further, another hypothesis is that sociophonetic features may be investigated by phonetic analyses, but they may be misleading in automatic ones.

5. *Methodology*

5.1 General methodology

A corpus was collected in order to reach the goals and check the hypotheses stated in § 4. The overall corpus considered here includes acoustic recordings of speech produced by 15 PD speakers (males, 8 from Bari and 7 from Lecce) and 10, though effectively 9[1], HC speakers (males, 5 from Bari and 4 from Lecce).

Participants were selected so as to be highly comparable in terms of age (PD: avg. 62,82, SD 8,17; HC: avg. 60,87, SD 8,75) and cognitive capabilities (Montreal Cognitive Assessment: PD=avg. 25, S.D. 1,1; HC= avg. 26,57, S.D. 1,61). HC reported no previous motor disorders and, consistently, showed a normal motor profile (Therapy Outcome Measure for Dysarthria – TOM-Dys: avg. 5, SD 0; Robertson's dysarthria Profile: > 134). PD participants varied greatly as for the number of years from diagnosis (avg.10,67, SD 6,62), but showed less variation as for their motor profile (TOM-Dys: avg. 3,71, SD 0,49; Robertson's Dysarthria Profile: <127).

Speakers were recorded reading aloud sentences included in 3 different short passages (14 sentences in total), sentences presented in isolation (9, produced with no specific expectation and attitude), and sentences preceded by a context that induced a specific, intended interpretation (16, including various modalities, such as declaratives and interrogatives, that are yes-no and wh-questions). Three repetitions were collected during recordings, which took place in a quiet room of the hospitals in Lecce and Bari.

The total amount of recorded speech considered here (excluding all pauses) is 56:75 minutes, 20:03 of PD dysarthric speech and 36:73 of HC speech.

Part of the corpus described above was used for fine-tuning and part for testing the automatic and phonetic evaluation. As for the former, the speech produced by 9 PD (5 from Bari and 4 from Lecce) and by 4 HC (2 from Bari and 2 from Lecce)

[1] Speech materials produced by one HC from Lecce were excluded due to technical issues.

was used; for the testing phase, the speech recorded by the remaining speakers was adopted, that is 6 PD speakers (3 from Bari and 3 from Lecce) and 5 HC (3 from Bari and 2 from Lecce). Speakers were assigned to the training or test set so as to have representative speakers in both sets and to have two groups as balanced as possible.

5.2 Methods for automatic analyses

In our study, we explored multiple CNN configurations. The configuration that achieved the best performance comprised four identical blocks, each containing a 2D convolutional layer, followed by batch normalization and an activation function. A spatial dropout layer was inserted between the first two layers to mitigate overfitting. Another configuration tested involved two blocks, each consisting of two stacked 2D convolutional layers, followed by ReLU activation, pooling, and dropout. Various hyperparameter adjustments were applied to optimize classification performance[2].

The automatic system was developed using a CNN trained in a cross-linguistic framework. Initially, the CNN was pre-trained on the PC-GITA database (Orozco-Arroyave, Arias-Londoño, Vargas-Bonilla, González-Rátiva & Nöth, 2014), a Spanish-language speech corpus designed for the analysis of individuals with PD. This dataset includes recordings from 50 PD patients and 50 HC speakers.

Following this pre-training phase, the model was fine-tuned using speech data from Italian speakers, which were excluded from the testing set to ensure unbiased evaluation. Due to the high level of noise in the Italian recordings, a preprocessing step was implemented to enhance audio quality. This involved the use of a speech enhancement system based on a Deep Neural Network (DNN), as proposed by Siniscalchi (2021), to improve signal clarity prior to PD detection.

The system was trained to perform a binary classification task, predicting the HC (0) or the PD (1) status of speakers who realized specific speech samples.

5.3 Methods for phonetic analyses

The analysis was carried out by means of a semi-automatic procedure. First, forced alignment was realized via MAUS (Schiel, 1999) trained on Italian, and then manual correction of segment boundaries and labelling of stop burst was performed in PRAAT (Boersma, 1992-2022). PRAAT scripts were then used to measure segment and Voice Onset Time (VOT) duration, as well as formant values (F1, F2), measured as the average value over the central 30 ms of vowels, for both vowels in stress and unstress syllables.

[2] For comparison, Support Vector Machines (SVMs) were also evaluated for classifying dysarthria severity. In our experiments, we employed standard SVM structures and fine-tuned the hyperparameters *C* and *gamma* within a range defined by powers of ten. Training was carried out using both linear and radial basis function (RBF) kernels to assess performance under different conditions. Related results are not discussed in this paper.

The analysis focussed on aspects that could highlight differences in accuracy in speech production and in the realization of sociophonetic traits. As for the accuracy in production, the analysis concerned the F1xF2 vowel space used for vowels /i/, /a/, /u/ in stress position, and the Vowel Articulation Index (VAI[3] – Roy, Nissen, Dromey & Sapir, 2009; Sapir, Ramig, Spielman & Fox, 2010; Skodda, Visser & Schlegel, 2011) to get information on the possible centralization of vowels, therefore again on the vowel space. Concerning consonants, the attention was focused on the accuracy in unvoiced stop production, analysed in terms of the number (percentage) of stops realized with a clear burst. Sociophonetic traits were analyzed by looking for differences across varieties and speakers both in the centralization of vowels /i/, /a/, /o/[4] in unstressed syllables (observed, again, by measuring the F1xF2 vowel space and the VAI), and in the aspiration of stops, by measuring differences in stop VOT duration (as a percentage of C duration).

Descriptive statistics, i.e. z-test of proportions, was used for analyzing percentages and T-test was used for analyzing VAI results.

6. *Results*

6.1 Automatic binary classification

Overall results obtained by means of the automatic evaluation in a binary classification task (HC vs. PD) are reported in the left part of Table 1, while accuracy results obtain for each subject are shown in the right part. Accuracy, sensitivity and specificity values range between 0.72 and 0.78, pointing to a satisfactory overall performance. Considering accuracy values for each subject, greater average accuracy is obtained for Bari PD and Lecce HC subjects. However, average values for Bari HCs and Lecce PDs are majorly affected by performances related to two subjects respectively (HC_BA_2 and PD_LE_8).

Table 1 - *Automatic evaluation in a binary classification task (HC vs. PD): Overall performance (left table) and by subject accuracy results (right table)*

Binary classification Overall	
Accuracy	0.758
precision	0.752
Recall	0.753
f1	0.752
roc_auc	0.753
sensitivity	0.784
specificity	0.722

Binary classification PD			*Binary classification HC*		
Speaker	Accuracy	Avg	Speaker	Accuracy	Avg
BA_3	1.000	0.96	BA_2	0.301	0.65
BA_5	0.907		BA_3	0.716	
BA_6	0.964		BA_5	0.927	
LE_4	0.860	0.66	LE_3	0.833	0.92
LE_6	0.836		LE_5	1.000	
LE_8	0.275				

[3] A lower VAI score points to a centralization of vowel formants.

[4] The vowel /o/ was chosen due to the lack of unstressed /u/.

The prediction of each participant belonging to the PD (1) or the HC (0) group was also calculated for each task (see Table 2). Results show that each speaker is quite accurately rated along a continuum ranging from HC to PD: on the one hand, PD subject productions received average evaluations greater that 0.6-0.7, besides speaker PD_LE_8 who was rather rated as a HC subject (0.2); on the other hand, rates associated to HC subjects are lower than 0.3-0.4, besides speaker HC_BA_2 who received an average evaluation around than 0.5.

As far as the task is concerned, besides PD_LE_8, PD productions seem to get worse predictions in sentences than in passages, while on the contrary HC productions in passages are always worse than those concerning sentences. Finally, in some cases, the rating of some subjects shows a relatively high difference depending on the task (SD around 0.2 for speakers PD_LE_4, HC_BA_3 and HC_LE_3).

Basically, PD productions were rated as HC in about 25.54 % of sentences, while HC realizations were rated as PD in about 19.14 % of sentences.

Table 2 - *Predictions of PD vs. HC status for each subject and each task*

		Sentences	*Passages*	*Average*	*SD*
PD	BA_3	1	1	*1*	0
	BA_5	1	1	*1*	0
	BA_6	*0.969*	0.929	*0.949*	0.028
	LE_4	0.634	*0.929*	*0.781*	0.208
	LE_6	*0.720*	0.643	*0.681*	0.055
	LE_8	*0.250*	0.143	*0.196*	0.076
HC	BA_2	0.515	*0.538*	*0.527*	0.016
	BA_3	0.208	*0.500*	*0.354*	0.206
	BA_5	0	*0.071*	*0.036*	0.051
	LE_3	0.125	*0.429*	*0.277*	0.215
	LE_5	0	0	*0*	0

As far as the impact of the variety is concerned, automatic rating of PD speech produced by Bari speakers tend to be slightly worse (closer to 1) than that produced by speakers from Lecce. On the other hand, the ratings of HC seem to be more balanced across the two varieties. In fact, excluding HC_BA_2 and PD_LE_8 who are particularly deviant and seem to correspond to borderline speakers, PDs in Bari and Lecce on average are rated 0,983 and 0,731, respectively; HCs in Bari and Lecce are rated 0,195 and 0,138 respectively.

6.2 Phonetic evaluation

The analysis of segmental characteristics in the whole set of test sentences rated automatically offers interesting results.

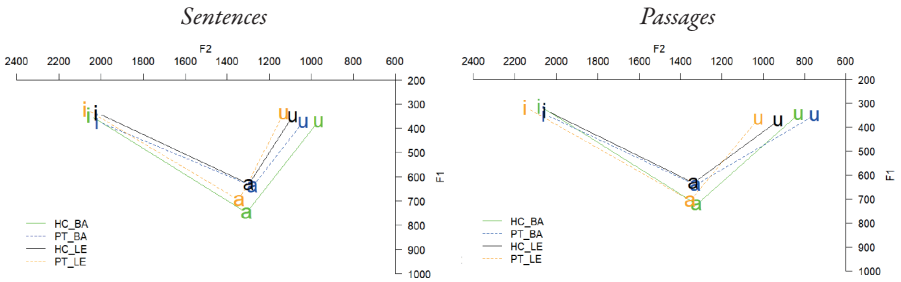
As for the production of stressed syllables, the VAI values show a slight centralization of vowels in PD productions in comparison to HC ones, though

differences are not significant either overall (Table 3, two upper rows) or when comparing speakers within variety (Table 3, bottom rows and Figure 1). However, plotting the raw Hz data, the vowel space seems to be reduced due to a decreased F1 value (reduced tongue depression) in the productions of /a/, at least by HC_LE and PD_BA speakers (Figure 1).

Table 3 - *Vowels/i/, /a/, /u/ in stressed syllable realized in sentences and passages by HCs and PDs from Bari and Lecce: VAI mean values and T-test (p-value)*

		Sentences	*T-test*	*Passages*	*T-test*
	HC	0,914	0,510	0,961	0.795
	PD	0,886		0.941	
BA	HC	0.948	0.242	1.0002	0.842
	PD	0.874		0.981	
LE	HC	0.862	0.494	0.901	0.989
	PD	0.898		0.899	

Figure 1 - *Vowels/i/, /a/, /u/ in stressed syllables realized in sentences (left) and passages (right) by HCs and PDs from Bari and Lecce (colors/lines): F1xF2 (Hertz) plots*

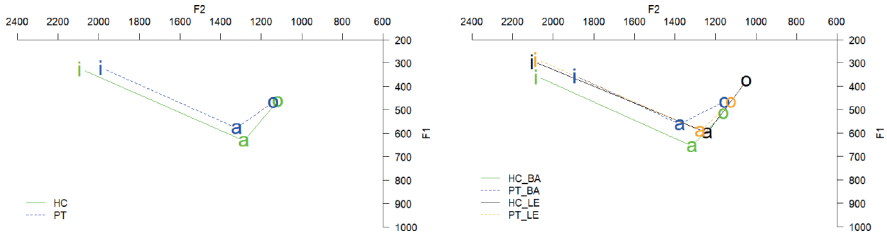


Concerning the realization of unstressed syllables, the VAI values show a slight centralization of vowels in PD productions in comparison to HC at least in sentences. The values do not change significantly (Table 4, two upper rows), besides a slightly significant difference for Bari speakers in sentences and Lecce speakers in passages (Table 4, bottom rows). Plotting the raw Hz data allows us to observe that the vowel space seems to be reduced due to a decreased F1 value (reduced tongue depression) in the productions of /a/ and a reduced F2 value (reduced advancement of the tongue) in /i/ (Figure 2, left); however, strategies may change depending on the variety, as the reduction may also concern the production of the /o/ vowel (e.g. Lecce PDs, Figure 2, right).

Table 4 - *Vowels/i/, /a/, /o/ in unstressed syllable realized in sentences and passages by HCs and PDs from Bari and Lecce: VAI mean values and T-test (p-value)*

		Sentences	*T-test*	*Passages*	*T-test*
	HC	0,851	0,113	0,783	0.624
	PD	0,784		0.794	
BA	HC	0.816	0.065	0.804	0.137
	PD	0.737		0.766	
LE	HC	0.904	0.285	0.751	0.058
	PD	0.831		0.822	

Figure 2 - *Vowels/i/, /a/, /o/ in unstressed syllables realized in sentences by all HCs and PDs (left) and by HCs and PDs from Bari and Lecce (right): F1xF2 (Hertz) plots*



The accuracy in plosive consonant production was checked by verifying the presence of a burst. Percentages of accurate productions are reported for HCs and PDs together with the results of z-test comparison (Table 5). Among the unvoiced plosives, PDs produced less accurately than HCs both bilabial stops (in both sentences and passages) and alveolar ones (only in sentence production), while no significant difference was found in the production of alveolars in passages and velars.

As far as the difference between Lecce and Bari speakers is concerned (see ** in Table 6), the only significant results concern the PD realization of /p/ in sentences, which is more accurate in Lecce PD speaker productions, and the /t/ realization in passages by HCs, which is again more accurate in the Lecce group.

Table 5 - */p/, /t/, /k/ production in sentences and passages by HCs and PDs: Accuracy of stop production as the percentage and total number of plosives and z-test comparison*

		Sentences	*z-test*	*Passages*	*z-test*
/p/	HC	100 %(320)	<0,0001	87,67 %(73)	<0.05
	PD	84,62 %(403)		69,32 %(88)	
/t/	HC	94,61 %(167)	<0,05	86,70 %(188)	n.s.
	PD	87,79 %(213)		84,75 %(236)	
/k/	HC	-	-	69,57 %(46)	n.s
	PD	-		65,22 %(69)	

Table 6 - *Accuracy in /p/, /t/, /k/ production in sentences and passages by HCs and PDs from Bari and Lecce (percentage and total number of plosives) and z-test comparison*

			Sentences	*z-test*	*Passages*	*z-test*
BA	/p/	HC	100 %(147)	<0,0001	92,86 %(42)	<0,05
		PD	74,4 %(168)**		68,18 %(44)	
	/t/	HC	93,75 %(80)	-	77,14 %(105)**	<0,05
		PD	88,64 %(88)		84,95 %(113)	
	/k/	HC	-	-	64 %(25)	-
		PD	-		72,5 %(40)	
LE	/p/	HC	100 %(173)	<0,0001	80,64 %(31)	-
		PD	91,91 %(235)**		70,45 %(44)	
	/t/	HC	95,40 %(87)	-	98,8 %(83)**	<0,05
		PD	87,2 %(125)		84,55 %(123)	
	/k/	HC	-	-	76,2 %(21)	-
		PD	-		55,17 %(29)	

7. *Discussion and conclusions*

The main goal of the work presented here was to observe if and how phonetic and automatic classification of dysarthric and healthy speech can differ on the same data. Further, a second aim was to check the role of sociophonetic difference in the speech data, in particular to understand if automatic analyses of speech may be deceived by sociophonetic traits that, when sociophonetics is not explicitly taken into account, could be considered as instances of inaccuracy in speech production. For the purpose of the analysis, a corpus was collected (56:75 minutes of articulated speech), including acoustic recordings of speech produced by 15 PD speakers (males, 8 from Bari and 7 from Lecce) and 10 HC speakers (males, 5 from Bari and 4 from Lecce, as one speaker was discarded for technical issue). Part of the corpus was used for fine-tuning the system for automatic analysis, and part for testing both the automatic and phonetic evaluation.

As for the overall performance of the automatic system, accuracy, sensitivity and specificity values range between 0.72 and 0.78, pointing to a satisfactory overall performance given the task and the data. Considering accuracy values for each subjects, greater average accuracy is obtained for Bari PD and Lecce HC subjects, with results that are clearly deviant in the case of two subjects, who are taken to be borderline (HC_BA_2 and PD_LE_8). The prediction of each participant belonging to the PD (1) or the HC (0) group turned out to rate speakers' production along a continuum ranging from HC to PD, even though results also depend on the task. In particular, PD productions seem to get worse predictions in sentences than in passages, while the opposite is true for HCs. The rating of subject productions along a continuum may easily correspond to different speaker performances; further, the different results concerning the task could correspond to an actual greater difficulty for PD in interpreting sentences than in reading

passages. In fact, sentences are shorter and clearly separated from each other, and the repetitive starts could possibly challenge PDs; further, in most cases required an interpretation that depended on the previous context. Finally, as far as the impact of the variety is concerned, the automatic rating of PD speech seems to be slightly worse in the case of Bari than Lecce speakers; however, the ratings of HCs seems to be more comparable across variety. The tendency may than be due to possible slight differences in the severity of dysarthria, as even though participants were chosen to represent a similar severity level (as measured by TOM-Dys and RDP; see § 5.1), Bari PDs could be slightly more compromised than the Lecce one. On the other hand, absolutely external factors, such as the quality of the recordings, seem to be not at play, as HC recordings took place in the same setting used for PDs and the result of their analysis seems to be comparable across varieties.

The phonetic analysis of vowels and stops produced in the all set of test sentences rated automatically was realized. As far as the production of vowels in stressed syllables is concerned, a slight centralization of vowels in PD production was found in comparison to HC realization, apparently mainly due to the reduced tongue lowering for /a/ (at least in Bari PDs and Lecce HCs), though PDs' VAI values were not significantly different from HCs' either overall or within variety. More interesting results in differentiating PDs and HCs are found in analyzing vowels realized in unstressed syllables: a slight centralization is observed for vowels in PD productions in comparison to HC at least in sentences for Bari speakers. However, the vowel space seems to be reduced not homogeneously, and the reduction does not necessarily involve the low vowel /a/ (e.g., the F1xF2 plots suggest that in sentence production, Lecce PDs seem to change the /o/ vowel aperture more than the /a/ vowel). As for the accuracy in unvoiced plosives, PDs produced less accurately than HCs both bilabial stops (in both sentences and passages) and alveolar ones (only in sentence production), while velars and alveolars in passages do not significantly differ. Finally, as for sociolinguistic differences the only significant results concern the PD realization of /p/ in sentences and that of /t/ by HCs in passages: in both cases, accuracy is greater in the Lecce group.

A comparison of the results of the automatic and the phonetic analysis aimed at distinguishing PD and HC subjects confirms our first hypothesis. The phonetic and automatic procedures may classify speech consistently, but do not allow for easy comparison: the automatic analysis rates speakers along a continuum ranging from HC to PD, even though results also depend on the task as PD productions seem to get worse predictions in sentences. On the other hand, the acoustic analysis focusing on specific segments and measures, does not consistently tease apart the two populations. Rather, it signals tendencies to reduce the vowel space, especially when vowels in unstressed syllables are considered, and produce unvoiced plosive consonants less accurately, especially bilabials and alveolars in sentences. Differences depending on the task are detected also in acoustics, as the results on alveolars show. Therefore, none of the two methods proved to be ideal. Rather, automatic and phonetic analysis could be successfully integrated depending on to goal of the

classification: if one also wants to analyze the reason for the results, phonetic analysis is to be preferred, whereas, if one wants to obtain a result very quickly, automatic analysis is, not surprisengly, to be preferred.

Regarding the second goal, that is understanding if automatic analyses of speech may be deceived by sociophonetic traits, results are not conclusive. The automatic rating of PD speech seems to be slightly worse in the case of Bari than Lecce speakers and, in theory, this could be related to differences in vowels (centralized in Bari, at least in unstress syllables) and consonant production (potentially considered as more accurate in Lecce due to the longer VOT expected for aspiration, requiring a clear burst). Consistently, also the phonetic analysis suggest that Lecce PD speakers are more accurate. However, a possible difference between the automatic and the phonetic approach may lie in the impact of the quality of the recordings: phonetic analyses, such as that presented here, are theory-driven and carried out semiautomatically, ensuring a systematic check of the data and speech phenomena, whereas a completely automatic treatment could be more easily affected by differences in the quality of recordings, that do not relate with speech production per se. Crucially, such difference does not mean that the automatic analysis was misled by sociophonetic traits. Further, if more data were available, the system could be trained on speech of a specific variety, making the system more robust as for sociophonetic features.

All in all, given the data considered here, automatic and phonetic analysis seem to offer similar results. However, they could be successfully integrated depending on to goal of analysis: if results are to be analyzed to get a better comprehension of the way speech is articulated, also taking into account sociophonetic features, the phonetic analysis is to be preferred; on the other hand, if a classification is to be achieved very quickly, with no specific interest in what speech characteristics are making the difference, the automatic analysis is, no surprise, to be preferred.

Acknowledgements

This work was funded by the PRIN 2017 project 2017JNKCYZ. We thank M.L. Fiorella and her collaborators, M. Bosco and D. Sciancalepore, for enrolling and recording experimental subjects in Bari; V. Sallustio, A. Trittola and R. Mitolo for subjects in Lecce. We warmly thank all the pathological and control subjects who participated in the experiment.

References

Audibert, N., Fougeron, C., Fredouille, C., Meunier, C. & Panseri, O. (2010). Evaluation d'un alignement automatique sur la parole dysarthrique. In *Mons: Actes des 28e Journées d'Etudes sur la Parole* (JEP'10), 353-356.

Bang, Y.-I., Min, K., Sohn, Y.H. & Cho, S.-R. (2013). Characteristics of vowel sounds in patients with Parkinson disease. In *NeuroRehabilitation*, 32(3), 649-54.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Best, C.T. (1995). A direct realist view of cross-language speech perception. In Strange, W. (a cura di), *Speech perception and linguistic experience: issues in cross-language research*. York Press, 171–204.

Browman, C.P., Goldstein, L. (1986). Articulatory gestures as phonological units. In *Phonology*, 6, 151-206.

Fussi, F., Cantagallo, A. (1999). *Profilo di valutazione della disartria*. Adattamento italiano del test di Robertson, raccolta di dati normativi e linee di trattamento, Ed. Omega.

Gili Fivela, B., d'Apolito, S. & Sigona, F. (2022a). Vowel space in hypokinetic dysartria: Preliminary investigations. In Dovetto, F.M., Raso, T. & Sorianello, P. (a cura di), *Le patologie del linguaggio: studi e risorse tra crossdisciplinarità e interdisciplinarità, Chimera (ISSN 2386-2629)*, 9, 147-164.

Gili Fivela, B., d'Apolito, S. & Pagliaro, A.C. (2022b). Tra economia dello sforzo e accuratezza nella disartria ipocinetica/ Between economy of effort and speech accuracy in hypokinetic dysartria. In Orrico, R., Schettino, L. (a cura di), *La posizione del parlante nell'interazione: atteggiamenti, intenzioni ed emozioni nella comunicazione verbale [The position of the speaker in interaction: attitudes, intentions, and emotions in verbal communication]*, *Studi AISV 9*. Milan: Officinaventuno, ISBN 978-88-97657-63-7, 99-114.

Gili Fivela, B., Pagliaro, A.C. & d'Apolito, S. (2023). Phonological and sociophonetic information in dysarthric speech: a first articulatory investigation on Italian. In Skarnitzl, R., Volín, J. (a cura di), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, 7-11 August 2023, Guarant International, 3937-3941.

Kim, H., Gurevich, N. (2021). Positional asymmetries in consonant production and intelligibility in dysarthric speech. In *Clinical Linguistics & Phonetics*, 37(2), 125-142.

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease (2003). State of the Art Review The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations. *Movement Disorders*, Vol. 18(7), 738–750.

Loporcaro, M. (1997). Puglia & Salento. In Maiden, M., Parry, M. (a cura di), The Dialects of Italy, London: Routledge, 338-348.

Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C. & Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Language Resources and Evaluation Conference*, (LREC), 2014, 342–347.

Robertson, S.J. (1982). Dysarthria profile. Background and development. In *College Speech Therapist Bullettin*, 359.3.82.

Rohlfs, G. (1949). *Historische Grammatik der Italienischen Sprache und ihrer Mundarten*. Vol. 1. Lautlehre. Bern: Francke, Grammatica storica dell'italiano e dei suoi dialetti (edizione italiana), Fonetica, Torino: Einaudi, 1966, 439-446.

Roland, V., Huet, K., Harmegnies, B., Piccaluga, M., Verhaegen, C. & Delvaux, V. (2023). Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease. In *Frontiers in Psychology*, 14.

Romano, A., Manco, F. (2004). Incidenza di fenomeni di riduzione vocalica nel parlato spontaneo a Bari e a Lecce. In Albano-Leoni, F., Cutugno, F., Pettorino, M. & Savy,

R. (a cura di), *Atti del Convegno Naz. Napoli*, "Il Parlato Italiano", 13-15 Febbraio 2003), Napoli: D'Auria (CD-ROM).

Roy, N., Nissen, S.L., Dromey, C. & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: evidence for vowel space expansion following manual circumlaryngeal therapy. In *Journal of Communication Disorders*, 42, 124-135.

Sapir, S., Ramig, L.O., Spielman, J.L. & Fox, C. (2010). Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech. Journal of Speech. In *Language, and Hearing Research*, 53, 114–25.

Schiel, F. (1999), Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS*, San Francisco, August 1999, 607-610.

Skodda, S., Visser, W. & Schlegel, U. (2011). Vowel Articulation in Parkinson's Disease. In *Journal of Voice*, 25, 467–72.

Skodda, S., Gronheit, W. & Schlegel, U. (2012). Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease. In *PLoS ONE*, vol.7 (2), 1–8.

Siniscalchi, S.M. (2021). Vector-to-Vector Regression via Distributional Loss for Speech Enhancement. In *IEEE Signal Processing Letters*, vol. 28, 254-258.

Tjaden, K. (2000). An acoustic study of coarticulation in dysarthric speakers with Parkinson disease. In *Journal of speech*, language, and hearing research, 43 (2000), 1466-1480.

Vásquez Correa, J.C., Arias, T., Orozco-Arroyave, J.R. & Nöth, E. (2018). A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease. In *Proceedings of Interspeech* 2018, 456-460, doi: 10.21437/Interspeech.2018-1988

Wang, Q., Fu, Y., Shao, B., Chang, L., Ren, K., Chen, Z. & Ling, Y. (2022). Early detection of Parkinson's disease from multiple signal speech: Based on Mandarin language dataset. In *Front Aging Neurosci*. 14: 1036588.

Wong, M.N., Murdoch, B.E., & Whelan, B.M. (2010). Kinematic analysis of lingual function in dysarthric speakers with Parkinson's disease: An electromagnetic articulograph study. In *Int. JS-LPathology*, vol. 12, 414–425.

Wong, M.N., Murdoch, B.E. & Whelan, B.M. (2011). Lingual Kinematics in Dysarthric and Nondysarthric Speakers with Parkinson's Disease. In *Parkinson's Disease*, 1-8.