INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Identifying distinctive acoustic and spectral features in Parkinson's disease

Yermiyahu Hauptman[1], Ruth Aloni-Lavi[1], Itshak Lapidot[1], Tanya Gurevich[2,4,5]
Yael Manor[2,3], Stav Naor [2], Noa Diamant[2], Irit Opher[1]

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel
[2]Movement Disorders Unit, Dept of Neurology, Tel-Aviv Sourasky Medical Center, Israel
[3]Ono Academic College, Faculty of Health Professions, Communication Sciences and Disorders Department, Kiryat Ono, Israel
[4]Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel,
[5]Sagol School of Neuroscience, Tel Aviv University, Israel

yermiyahuh@afeka.ac.il, rutil@afeka.ac.il, itshakl@afeka.ac.il, tanyag@tlvmc.gov.il, yaelm@tlvmc.gov.il, stavnaor11@gmail.com, diamant.noa@gmail.com, opher.irit@gmail.com

Abstract

In this paper we try to identify spectral and acoustic features that are distinctive of Parkinson's disease patients' speech. We investigate the contribution of several features' families to a simple classification task that distinguishes between two balanced groups – patients with Parkinson's disease and their age and gender matched group of Healthy Controls, both uttering sustained vowels. We achieve over 75% correct classification using a combination of acoustic and spectral features. We show that combining a few statistical functionals of these features yields very good results.. This can be explained by two reasons: the first is that the statistics of Parkinson's disease patients' speech defer from those of Healthy people's speech; the second and more important one is the gradual nature of the Parkinsonian speech that is manifested by the changes within an utterance. We speculate that the feature families that most contribute to the classification task are the most distinctive for detecting the disease and suggest testing this hypothesis by performing long-term analysis of both patient and healthy control subjects. Similar accuracy is obtained when analyzing spontaneous speech where each utterance is represented by a single normalized i-vector.

Index Terms: Acoustic Features, Spectral Features, i-vectors, Parkinson's Disease,

1. Introduction

Parkinson's disease (PD) is a chronic, progressive neurodegenerative and multi- dimensional disease that involves range of disabling motor and non-motor symptoms [1]. PD affects 1-2 per 1000 of the population at any time. PD prevalence is increasing with age and PD affects 1% of the population above 60 years [2].

Speech dysfunction is one of the most common symptoms of PD. Speech disorders occur in 70%-90% of patients with PD during the course of the disease, with higher prevalence in advanced stages [3], [4]. Speech disorders associated with PD are known as "hypokinetic dysarthria" and the symptoms involve dysfunctions of several systems: respiration, phonation, articulation, resonance and prosody [5]. The ensuing communication disability affects not only the patients' quality of life but also imposes a burden on their families and surrounding contacts [4]. Overcoming this impairment remains a challenge, since speech disturbances are only partially responsive to medication therapy and may even worsen by deep brain stimulation [3] and [4].

Nowadays, in the era of advanced technologies, early detection of speech and voice disturbances becomes feasible. This may lead to early diagnosis of PD and initiate early treatment which may help maintain and improve communication abilities and quality of life of the patients. Furthermore, since most people who develop PD are elderly, it is important to develop technologies that can issue warnings based on daily or weekly assessments. Such technologies can also assist physicians, caretakers and patients to monitor the disease progression in a remote and non-invasive manner. This can be done using smartphones applications that can easily monitor speech [6] and movements [7].

When considering such tools, an important question to address is what features to use when analyzing and monitoring PD patients' speech, as various features are available. Some authors use standard MFCCs [8], while others employ additional features such as PLP-RASTA [9]. Another option is to use custom features and measures that are frequently used by speech therapists and neurologists, e.g. shimmer, jitter and NHR as in [10]. It is also possible to use multiple ASR feature combinations that are available in OpenSmile that was used in [11], or to define specific feature sets, e.g. the phonation feature set suggested in [12] that use various mathematical and statistical operations applied to F0, jitter, shimmer, NHR and HNR.

This research aims at detecting the most relevant features for characterizing the disorders in speech of PD patients, using a corpus of 70 PD patients and 70 Healthy Controls (HC) of the same age and gender distribution. The study's population includes PD patients mostly from medium and advanced stages. One of the limitations of this study is lack of analyses from early diagnosed PD patients.

Our motivation is to identify a small set of distinctive features that can be used at later phases mostly for long term

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2465

monitoring. To achieve this goal, we test various feature families separately by training a linear SVM classifier over vectors of the same feature families and then choosing the feature family that depicts the highest classification score. Next, we employ a greedy bottom up approach where we evaluate the contribution of couples and triplets of additional feature families to the classification task.

Parkinson speech is characterized by major changes throughout an utterance [13]. This behavior is reflected partially in fluctuations of the acoustic features throughout an utterance. Taking this into account, we employ several statistical measures that are calculated over the whole utterance such as Mean, Median and STD.

In addition, we observe that there is a gradual change in the speech amplitude during production of maximal phonation time (MPT). These changes occur for both PD and HC but are more prominent for PD patients probably since PD voice production requires a lot of effort, so it becomes weaker after a few seconds. We also notice spectral changes throughout MPT of PD patients. Therefore, we add minimal and maximal values and Max/Min ratio and some other statistical measures that can represent the variability and irregularity along a speech utterance. These measures are described in section 3.2 and 4 and constitute an important part of our analysis.

Since most of our analysis is done using MPT recordings, we wish to also compare the classification ability when analyzing longer and more variable utterances, Hence, we use normalized i-vectors that are extracted from all parts of our corpus, including the spontaneous speech task. [14].

The rest of the paper is as follows: in section 2 we describe the speech corpus we collected. Features are presented in section 3 and statistic parameters are explained in section 4. The classifier is described in section 5, followed by the evaluation protocol in section 6. Next the experiments and results are described in section 7. Conclusions and future work are presented at the end.

2. Speech Corpus

Our corpus includes six speech tasks for each subject:

1. Sustained vowel /ah/ for maximal phonation time (MPT).
2. Counting task.
3. A few repetitions of /PATAKA/.
4. A phonetically balanced reading paragraph in Hebrew (The Thousand Islands, [15]).
5. Picture description.
6. Free speech that is elicited by asking the subjects to speak about themselves, their family, their occupation or some recent personal experience.

The design of the speech corpus takes into account the difficulty of some PD patients to elicit speech, so picture description and spontaneous speech are the last two tasks.

Two groups of speakers were recorded. The first consists of PD patients and the second consists of matched HC. Both groups have 60% males and 40% females, and the average age of the subjects in both groups is 64 with a standard deviation of 10.6 for the PD group and 10.3 for the HC. Average time that passed since diagnosis is 10.8 years, with 7 years standard deviation. To the best of our knowledge, this is one of the largest speech corpora for PD research that is well balanced, and the only such corpus in Hebrew. Moreover, this is one of the only corpora that is accompanied with cognitive screening, subjective vocal assessment by the examiners and participants' report of the effect of the vocal functions on their everyday life. These measurements are, relevant for long term monitoring of PD progress (see future work in section 8).

All speakers underwent three standard cognitive assessment tests: MoCA [16], GRBAS [17] and VHI [18] before each recording session. Most subjects were recorded once, while 12 speakers of PD group were recorded a few times throughout a year, where at least 3 months passed between each consecutive recording. The long-term recordings are not included in the analysis described in this paper.

Data was recorded at 44.1kHz , 16bits per sample using the same microphone. No down sampling was done, and all features were extracted from the original records.

3. Features' description

Two different "families" of features are extracted. The first are i-vectors; The second consists of a variety of statistics that are calculated on different acoustic features. We begin by describing the features and then we discuss the classifiers.

3.1. I-vectors

For each part of the database, we conduct its own experiment. For each recording we extract a widely used in speaker recognition, i-vectors [19]. It is a length independent, fixed size representation. We use the same i-vector extractor as in [14]. To extract the i-vectors, first a *Gaussian mixture model-Universal background model* (GMM-UBM) has to be trained and then a *total variability* (TV) matrix. For this, we use Mel frequency cepstral coefficients (MFCC), that are extracted using a 25ms Hamming window with a frame rate of 100 frames per second. 19 MFCC features together with log-energy. Cepstral mean subtraction and variance normalization are applied to the MFCCs. These vectors are augmented by the delta and delta delta to produce 60-dimensional feature vectors. Male only UBM of 2048 Gaussians mixture components is trained using Fisher Part 1; Switchboard II, Phase 2; switchboard Cellular, Parts 1 and 2; and NIST 2004-2006 SREs. Then, the total variability matrix with a low rank of 400 is trained using labeled data from same databases as for the UBM. In total, 975 unique male speakers with 10705 sessions are used.

Two versions are examined: i-vectors and normalized i-vectors such that each i-vector has a norm that equals 1.The i-vectors are used as an input to a linear SVM classifier.

3.2. Other features

Most of the features we extract are standard, but some the statistics we extract based on those features are new (as described in section 4). These features and statistics are optimized for the evaluation over the MPT subset. The most common features are *mel-frequency cepstral coefficients* (MFCC), *mef filter bank* (MFB), *perceptual linear predictive* ***rela****tive* ***spec****t****ra****l* (PLP-RASTA) and *linear prediction coefficients* (LPC). In addition, we also extract time domain entropy (ET), spectral entropy (ES), LPC after Daubechies 3 wavelet filtering (DB3_LPC) and different statistics from discrete wavelet transform (DWT_Stat).

For all features, before extracting, three steps are taken (excluding ET, that skips the 3rd step):

1. Normalizing the speech signal by its *root mean square* (RMS),
2. Dividing the signal into 60 msec frames with a sliding window of 20msec .
3. Each frame is multiplied by a Hamming window.

We choose to work with 60 msec frames since PD speech is sometimes slower, so stationarity of speech is maintained for longer periods than those of HC. Nevertheless, for the MPT case, the signal is monotonic so should be stationary for longer duration for both, PD and HC speakers.

Next, we relate to the specific features where we focus on the less common ones.

3.2.1. Commonly used features

We will not explain the commonly used features, and only mention the dimensionality of each feature vector:

- MFCC – 17
- MFB – 16
- LPC – 13
- PLP-RASTA – 15

3.2.2. Time entropy features

For each frame of the speech signal (waveform) $s_k(n)$ a histogram with 10 bins is calculated and normalized by the number of samples in a frame, to obtain the *probability mass function* (PMF), $p_k(m)$. Then the entropy calculation is $H_T(k) = -\sum_{m=1}^{10} p_k(m) \log\{p_k(m)\}$. This is a one-dimensional feature.

3.2.3. Spectral entropy features

The entropy calculation is the same as in 3.2.2, but instead of calculating the entropy over the time signal, a periodogram of the frame is calculated to obtain $S_k(f)$. The ES feature is also one-dimensional, $H_S(k)$.

3.2.4. Daubechies 3 discrete wavelet filtering - LPC

This feature vector is based on [20]. The idea is to extract the LPC features, but not from the signal itself. Instead, for each speech frame $s_k(n)$, first the wavelet transform is performed (we use DB3 wavelets), and for each wavelet filter, $\psi_q(n)$ $q = 1,\ldots,Q$ an output signal is calculated, $s_k(q,n)$. For each output signal, LPC feature vector of order L is calculated. The final feature vector is of size $Q \times L$. In our case $Q = 5$ and $L = 17$, so the feature vector has dimensionality of 85 .

3.2.5. DWT_Stat features

When performing the wavelet transform, the output of each filter is a time signal $s_k(q,n)$. From each signal, different features can be extracted. The features we use are a subset of the features presented in [21]. Three features are calculated only for $q = 2,\ldots,8$. In total, 21 features per frame:

1. Mean of the $|s_k(q,n)|$,
2. Maximum value of the periodogram of $s_k(q,n)$: $\max\{S_k(q,f)\}$,
3. Be M_q number of frequency bin of $S_k(q,f)$. Be the expression $\sum_{m=1}^{\mu_q} f_m \cdot S_k(q,f_m) = 0.5 \cdot \sum_{m=1}^{M_q} f_m \cdot S_k(q,f_m)$; Then f_{μ_q} is the feature.

4. Statistical parameters

As mentioned in section 3.2, the features are extracted every 20msec . It means that for each recording, we might have several hundreds of feature vectors and more important is that this number varies from one recording to another. I-vector is one option to avoid this variability and to have a single length representation per recording. Here we present an additional approach, by calculating 10 different statistics over all feature vectors per recording. It means that final recording representation is $10 \times \#(Feature\ Vector)$, while # means the length. Having K frames at the recording, the statistics are:

1. For each dimension (feature) in the vector, an entropy is computed over the K values, using 10 bins PMF as in 3.2.2.
2. The mean value of each dimension.
3. The minimum value of each dimension.
4. The maximum value of each dimension.
5. The standard deviation of each dimension.
6. Slope of linear curve fitting of each coefficients over the entire recording.
7. Maximum value minus minimum value per dimension (statistics 4 minus statistics 3).
8. Taking the mean of the first 20% of the frames and the mean of the last 20%. The ratio per dimension is the feature.
9. For each feature type we perform *vector quantization* (VQ) of order 16, and the farthest feature from the VQ is the **outlier** statistic. Farthest means the maximal Euclidian distance between the feature vector to the closest code-word in the codebook.
10. Performing VQ of size 4 and taking the distance from the outlier vector. Similar to the way it is done in statistic 9.

As patients with Parkinson's disease are assumed not to hold stable the MPT task, we assume that these statistics can capture some irregularities in their speech. Statistics 9 and 10 are dedicated to capturing the outlier effects, which are more probable for patients with Parkinson's disease. The set of statistics and the parameters were designed by trial and error using a larger un-balanced speech corpus. It is true particularly for VQ sizes in 9 and 10, and the choice of 20% in 8.

5. The classifier

Several classifiers are examined during the research. The most stable classifier that works well for both i-vectors and statistic vectors is the *support vector machine* (SVM) with a linear kernel. All our results are presented for this classifier.

6. Evaluation protocol

Due to the scarcity of data, all experiments are conducted by a 5-fold evaluation protocol. All the data is divided into 5 folds, 20% of the data each (20% of the patients with Parkinson's and 20% of Healthy Controls). Each time, one fold is taken out; 4

folds are used for training the SVM, that is tested on the fold that is held out. This procedure is repeated 5 times, ones for each fold.

7. Experiments and results

First, we test the i-vectors, normalized and non-normalized. We test them over all speech tasks of our dataset, and summarize in terms of success rate [%] in Table 1. As expected, the normalized i-vectors perform better than the non-normalized for most tasks. Only in the MPT case, the results for the non-normalized i-vectors are better.

Next, we test all individual feature families (with statistics) and summarize the results in Table 2. These experiments are carried out using the MPT task that is the easiest to compare as it consists of a single vowel and allows close examination of differences between the two groups.

Table 1: *i-vectors vs. normalized i-vectors.*

Speech Task	i-vectors	norm. i-vectors
MPT	76.0	71.7
Counting	67.4	69.3
PATAKA	59.6	74.1
Reading	71.3	72.1
Picture	71.5	73.9
Free speech	70.5	74.1

As could be expected, the weakest statistics are from 1-dimensional feature vectors, time- and spectral-entropies. The time-entropy based classifier is a very weak classifier, almost as tossing a coin. The best classifiers are based on MFB and PLP-RASTA statistics. This is consistent with the findings in [9] where PLP-RASTA were found to perform better than MFCC or LPC. PLP-RASTA-statistics vector achieve slightly better results than the normalized i-vectors, but still lower than the non-normalized i-vectors for the MPT task. We need to remember that i-vectors are of dimensionality 400, while MFB-statistics has 160 dimensions and PLP-RASTA-statistics has 150 dimensions. Therefore, in the next experiment we use a greedy search to augment these two vectors (each time a different family) with other statistics' vectors. In Table 3 we summarized the best combinations.

Table 2: *MPT test evaluation on different feature families.*

Feature Family	Success rate [%]
MFCC	66.6
MFB	**68.0**
LPC	59.8
PLP-RASTA	**73.4**
Time Entropy	50.1
Spectral Entropy	57.6
DB3_LPC	65.9
DWT_Stat	64.2

Adding DB3_LPC to the MFB statistics increase the vector dimension to 1010, and with MFCC statistics to 1180 dimensions. This high cost in dimensionality, leads to more than 15% relative improvement in the success rate. For PLP-RASTA we find that the best couple is MFB, however it does not yield any improvement. Surprisingly, adding one dimensional spectral entropy (in total 320 dimensions), leads to a nice improvement, identical to adding LPC (in total 460 dimensions).

Table 3: *Augmentation of MFB and PLP-RASTA.*

Feature Combination	Success rate [%]
MFB	66.6
MFB+ DB3_LPC	74.8
MFB+ DB3_LPC+MFCC	**76.9**
PLP-RASTA	73.4
PLP-RASTA+MFB	73.2
PLP-RASTA+MFB+ES/LPC	**77.2**

8. Conclusions

In this paper we explore separability power of different feature families in distinguishing between PD patients and HC.

This allows us to pinpoint specific feature families that can be used to analyze and assess PD speech. In addition, we see that i-vectors that capture the whole recording, achieve good separation as well, which is promising for spontaneous speech analysis.

Next, we try to see if there are other global statistics that can provide information regarding the entire recording, similar to the i-vectors. Several options are examined, based on standard features, such as MFCCs, and less standard, such as Spectral Entropy and DB3_LPC. We find that several combinations can outperform i-vectors results, mostly relying on commonly used features, except for the nonstandard ES. This good separation ability can be attributed to the various statistical measures that capture the irregularity in PD speech that is manifested in changes throughout a speech utterance, as opposed to the more regular speech of HC. Our results show that PD affected voice and speech, as dependent on laryngeal, respiratory and articulatory functions, can still be well captured and represented by Mel based analysis. The variations that are more evident in Parkinsonian speech, are well described by various statistics of standard features. This means that traditional speech features can carry highly relevant information for PD speech, as they do for HC, while the distributions of these features change for the two groups. Rare events for healthy speakers become more probable for PD speakers, e.g. soft and breathy voice or decreased speech rate. This behavior is reflected in the separation ability of the statistics of the standard feature families.

A lot of room is left for further investigation. Other specialized statistics can be extracted. Standard manipulations on i-vectors such as mean subtraction or *principle component analysis* (PCA) can be tested, but the size of the corpus makes it difficult, so this can be done only in the nested cross-validation on the training dataset. We intend to test it is the next phase, and apply it to the new presented statistics as well. Another planned expansion of this work relates to long term monitoring. We plan to analyze small changes in the values of the most distinctive feature families along the long-term recordings and evaluate correlation between these changes and the objective repetitive clinical evaluations of patients and HC.

9. Acknowledgements

This work was partially done with the support of research grant no. 61045 by the Israel Innovation Authority.

10. References

[1] N. Giladi, Y. Manor, A. Hilel, and T. Gurevich, "Interdisciplinary teamwork for the treatment of people with Parkinson's disease and their families," *Current Neurology and Neuroscience Reports*, vol. 14, no. 11, pp.493, November 2014.

[2] O.B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease", *J. of Neural Transm*, vol. 124, no. 8, pp. 901-905, Aug. 2017

[3] G. M. Schultz and M. K. Grant, "Effects of speech therapy and pharmacological and surgical treatments on voice and speech in Parkinson's disease: A review of the literature," *J. Comm. Disorders*, vol. 33, no. 1, pp. 59-99, February 2000.

[4] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *J. Speech Hearing Disord.*, vol. 43, no. 1, pp. 47-57, Feb. 1978.

[5] S. Sapir, L. Ramig, and C. Fox, "Speech and swallowing disorders in Parkinson's disease," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 16, no. 3, pp. 205-210, June 2008.

[6] P. Klumpp, T. Janu, T. Arias-Vergara, J. C. Vasquez Correa, J. R. Orozco-Arroyave and E. Noth, "Apkinson – A Mobile Monitoring Solution for Parkinson's Disease", *Interspeech'17*, August 20-24, 2017, Stockholm, Sweden.

[7] C. Stamate, G. D.,Magoulas, S. Kueppers, E. Nomikou, I. Daskalopoulos, M. U. Luchini, T. Moussouri, and G. Roussos, "Deep learning Parkinson's from smartphone data", *IEEE Int. Conference on Pervasive Computing and Communications*, 2017.

[8] P. Schwab and W. Karlen, "PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data", *AAAI-2019*, Jan. 27-Feb. 1, 2019, Honolulu, Hawaii,USA.

[9] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease," *Applied Soft Computing*, vol. 62, pp. 649-666, 2018.

[10] L. F. Silva, A. C. Gama, F. E. Cardoso, C. A. Reis, and I. B. Bassi, "Idiopathic Parkinson's disease: vocal and quality of life analysis," *Arq Neuropsiquiatr*, vol. 70, pp. 674-679, 2012.

[11] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully Automated Assessment of the Severity of Parkinson's from Speech," *Computer Speech and Language*, vol. 29, no.1, pp. 172-185, Jan. 2015.

[12] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease". *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[13] K. Tjaden, "Speech and swallowing in Parkinson's disease," *Top Geriatr Rehabil*, vol. 24, no. 2, pp 115-126, 2008.

[14] I. Salmun, I. Opher, I. Lapidot, "On the Use of PLDA i-vector Scoring for Clustering Short Segments," *Proceedings of Speaker Odyssey 2016*, June 21-24, 2016, Bilbao, Spain.

[15] O. Amir. "Thousand Islands - Hebrew reading passage: preparation and validation". 10.13140/RG.2.1.2837.0809, 2006.

[16] K. E. McKee and M. E. Hackney, "The four square step test in individuals with Parkinson's disease: Association with executive function and comparison with older adults," *Neurorehabilitation*, vol. 35, no.2, 2014.

[17] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, and C. Croux, "Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality", *Official journal of the Voice Foundation*, vol. 11, no. 1, pp 74-80, 1997.

[18] V. N. Young, L. J. Smith, and C. Rosen, "Voice outcome following acute unilateral vocal fold paralysis," *The annals of otology, rhinology and laryngology*, vol. 122, no. 3, pp. 197-204, 2013.

[19] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol 19, no. 4, pp. 788-798, May 2011.

[20] N. S Nehe and R. S Holambe, "DWT and LPC based feature extraction methods for isolated word recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, January, 2012.

[21] A. Phinyomark, A. Nuidod, P. Phukpattaranont, and C. Limsakul "Feature extraction and reduction of wavelet transform coefficients for EMG pattern classification," *Elektronika ir Elektrotechnika* (*Electronics and Electrical Engineering)*, vol. 122, no. 6, pp. 27-32, 2012.