Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients

Laetitia Jeancolas, Habib Benali, Badr-Eddine Benkelfat, Graziella Mangone, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehericy, Dijana Petrovska-Delacrétaz

Automatic Detection of Early Stages of Parkinson's Disease through Acoustic Voice Analysis with Mel-Frequency Cepstral Coefficients

Laetitia Jeancolas, Habib Benali, Badr-Eddine Benkelfat, Graziella Mangone, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehéricy, Dijana Petrovska-Delacrétaz

3rd International Conference on Advanced Technologies
for Signal and Image Processing - ATSIP'2017,
May 22-24, 2017, Fez, Morroco.

Automatic Detection of Early Stages of Parkinson's Disease through Acoustic Voice Analysis with Mel-Frequency Cepstral Coefficients

Laetitia Jeancolas*, Habib Benali†‡, Badr-Eddine Benkelfat*, Graziella Mangone§, Jean-Christophe Corvol§,
Marie Vidailhet§, Stephane Lehericy§ and Dijana Petrovska-Delacrétaz*
*SAMOVAR, UMR 5157, Télécom SudParis,
CNRS, Université Paris-Saclay, France
Email: {laetitia.jeancolas, badr-eddine.benkelfat, dijana.petrovska}@telecom-sudparis.eu
† PERFORM Centre, Electrical & Computer Engineering Department
Concordia University, Montreal, QC, Canada
‡ Laboratoire d'Imagerie Biomédicale INSERM U1146 - CNRS UMR 7371-UPMC UM CR2
Pierre & Marie Curie University, Paris, France. Email: habib.benali@lib.upmc.fr
§ Sorbonne University, UPMC Univ Paris 06 UMR S 1127; INSERM U 1127 and CIC 1422; CNRS UMR 7225;
Brain and Spine Institute ICM and Centre de Neuroimagerie de Recherche (CENIR);
Assistance Publique Hôpitaux de Paris APHP, Department of Neurology, Hôpital Pitié-Salpêtrière, F-75013, Paris, France.
Email: graziella.mangone@icm-institute.org; {jean-christophe.corvol, marie.vidailhet}@aphp.fr; stephane.lehericy@upmc.fr

***Abstract*—Vocal impairments are one of the earliest disrupted modalities in Parkinson's disease (PD). Most of the studies whose aim was to detect Parkinson's disease through acoustic analysis use global parameters. In the meantime, in speaker and speech recognition, analyses are carried out by short-term parameters, and more precisely by Mel-Frequency Cepstral Coefficients (MFCC), combined with Gaussian Mixture Models (GMM). This paper presents an adaptation of the classical methodology used in speaker recognition to the detection of early stages of Parkinson's disease. Automatic analyses were performed during 4 tasks: sustained vowels, fast syllable repetitions, free speech and reading. Men and women were considered separately in order to improve the classification performance. Leave one subject out cross validation exhibits accuracies ranging from 60% to 91% depending on the speech task and on the gender. Best performances are reached during the reading task (91% for men). This accuracy, obtained with a simple and fast methodology, is in line with the best classification results in early PD detection found in literature, obtained with more complex methods.**

***Keywords*—Parkinson's disease; automatic detection; mfcc; speech signal processing; classification**

I. Introduction

Parkinson's disease (PD) is a neurodegenerative disease whose prevalence increases with age. It affects 1% of people older than 60 years, and up to 4% of those over 80 [1]. It is the second most common neurodegenerative disease after Alzheimer's disease. Symptoms are mainly motor symptoms which result from a loss of dopaminergic neurons in the substantia nigra (a structure located in the midbrain area), and disrupted connections between substantia nigra and basal ganglia. The standard diagnosis relies on motor tests that are positive when they reveal two out of three of the following symptoms: akinesia (difficulty to start a movement), rigidity and rest tremor. Unfortunately, those motor symptoms appear once 70% of dopaminergic neurons are destroyed. For this reason, the detection of Parkinson's disease at an earlier stage, is one of the main goals of PD research.

One method that has the potential to detect early changes in PD patients is the acoustic analysis of voice. Indeed, people with Parkinson's disease have vocal impairments characterized by hypokinetic dysarthria. This includes perturbation in prosody, articulation and phonation. Their voice is more monotonous (with a diminution of intensity and pitch modulations). Speech flow is altered and patients make more dysfluencies. Consonant articulation is imprecise, particularly for the occlusive consonants (/p/, /t/, /k/, /b/, /d/, /g/). Vowel articulation is also impaired: differences between vowels tend to decrease, which results in a reduction of the vowel surface area. As for phonation, pitch and intensity are unsteady (particularly during sustained vowel tasks), and timbre is hoarse. More details about those vocal impairments can be found in [2].

One of the main interests of voice analysis in PD is that vocal impairments are present from the beginning of the disease and even several years before a clinical diagnostic can be made [3]–[5]. Moreover automatic detection of PD based on voice acoustic analysis reaches an accuracy of 95% [6]. As for specific detection of the early and middle stages of PD, the best accuracies are around 90% [5]–[7]. Most common speech tasks for vocal analyses in PD detection are the following:
- Sustained vowels: vowel /a/ is the most common. This task reveals phonatory impairments.
- Diadochokinesia (DDK) task: fast syllables repetition, usually syllable with occlusive consonant (like /pa/-/ta/-/ka/). This task reveals the consonant articulation impairments.
- Reading and free speech: to analyze consonant and vowel articulation as well as prosody.

Initial studies performed long term signal analyses. They extracted global features such as number of pauses, number of dysfluent words, Standard Deviation (SD) of pitch and of

intensity. They also averaged low-level perturbations, such as shimmer, jitter, voice onset time, signal to noise ratio, formants or vowel space area, in order to have one parameter per subject and per task. Once the features were extracted, they performed significance and redundancy tests to select a reduced set of features. Selected features were then fed as inputs to classifiers (Support Vector Machine (SVM) was the most commonly used). Finally, most of the time, in order to have a reliable estimation of the accuracy, they validated their classifier with cross validation. Leave One Subject Out (LOSO) and 10-fold cross validation were the two cross validation techniques the most used in this context.

Most of those long term acoustic parameters require an accurate estimation of the fundamental frequency, which is a hard task, above all for pathological voices. On the other hand, other parameters are widely used in speaker and speech recognition: Mel-Frequency Cepstral Coefficients (MFCC), associated with Gaussian Mixtures Models (GMM). MFCC are short-term parameters: they are calculated on short time windows (between 20ms to 40ms). They characterize the spectral envelope on a Mel scale (reflecting the natural auditory perception). These features have the advantage not to rely on pitch estimation. For a decade, MFCC have been used for the detection of different voice impairments, like dysphonia [8]–[10]. They were introduced for PD disease detection a few years later, the first time by Tsanas in 2012 [11], and later by several other studies [12]–[18]. These authors extracted some statistics from the MFCC (mean, and in some cases: SD, kurtosis and skewness), sometimes in addition with other classical features, and fed them to their classifier. They calculated statistics from the whole utterance when the task was acoustically steady (like sustained vowels) [11], [15]–[18], or on portions of the tasks that shared some common acoustical characteristics, like voiced and unvoiced frames within a word [12]. The accuracies achieved with the sustained vowel tasks ranged from 80% to 90%. Tsanas [11] and Jafari [17] obtained 99% and 97.5% accuracies, but their validation process was not speaker-independent (the utterances used in the training groups were different than the ones used for the test but could belong to the same subject). This can lead to optimistic and biased accuracies. Authors from [12], who performed segmentation before extracting MFCCs and other parameters (on voiced and unvoiced frames), reported accuracies for PD detection (not especially early PD) between 84% to 99%.

In order to get information on frames that are acoustically very different (for example if we take all the frames from a reading task) we have to model more precisely the MFCC distribution. One possible way chosen in some studies [19], [20] is to use Vector Quantization (VQ). In [20], one codebook was computed per subject. The best accuracy obtained after a LOSO cross validation (on a SVM) was 82%. In [19] a calculation of VQ distortion between test subjects and training subjects enabled a classification for different codebook sizes. The best performance (93%) was obtained for a codebook size of 16. Another method to model more precisely the MFCC distribution is to use multi-dimensional Gaussian Mixture Models (GMM) as an approximation of the multi-dimensional probability density function of MFCC. For speech features it has been showed that a finite number of Gaussians is sufficient to form a smooth approximation of the probability density function [21]. Bocklet et al. [22] in 2013 used this method to model the MFCC distribution during different types of speech tasks (sustained vowels, free speech, reading, DDK). They created one multi-dimensional GMM (with 128 Gaussians) per subject and per task with Universal Background GMM modeling, and kept the means of the 128 Gaussians. These multi-dimensional features vectors (one per task and per speaker) were then used as inputs for the SVM classifier. The best accuracy they obtained was 81%, reached during a reading task. This method was also tested by [6] for three different languages during reading tasks (for which they obtained an accuracy around 80%) and during DDK task (with accuracies from 70% to 87% depending on the language).

In speaker recognition, usually men and women are treated with separate classifiers. Moreover it has already been shown that differentiation by gender increased performance in detection of some voice impairments, like laryngeal pathologies [23] when MFCC are used. We decided to follow this direction and see how we can exploit the GMM modeling in order to distinguish early stages of PD from healthy controls. Unlike the previous studies which used GMM as a step in their features extraction [6], [22], we employed them as a component of our classifier. We created one GMM per class and merely calculated the likelihood of the test subjects' MFCC against the two GMM models (one for PD class and one for control class). No SVM or other further classifier was then used. We wanted to see if we could have a good classification performance for early stages of PD detection with this simple, automatic and fast method.

II. Voice Acquisition

A. Participants

PD patients and healthy controls were recruited at the Pitié-Salpêtrière Hospital in the context of a longitudinal study (ICEBERG). Additional healthy control subjects were recruited via an offer of the RISC (French information relay in cognitive science). The inclusion criteria for PD patients were: a diagnosis of idiopathic Parkinson's disease, according to the UK Parkinson's Disease Society Brain Bank (UKPDSBB) clinical diagnostic criteria [24]; less than 4 years of disease duration at inclusion; dopamine transporter deficiency detected with a DATScan; absence of atypical Parkinsonian syndrome (like multiple system atrophy, Lewy body dementia, progressive supranuclear palsy); no Parkinsonian syndrome due to neuroleptic or MPTP neurotoxin. The inclusion criteria for control subjects were: having a normal neurologic examination and being age-matched with the patients' groups (between 40 and 70 years old). All subjects (PD and controls) had a medical examination, motor and cognitive tests, biological sampling and medical imaging (DATScan, MRI and fMRI).

For our study we recorded 74 French subjects among these participants. 40 were recently diagnosed with PD (21 males and 19 females). 34 were healthy control subjects (14 males and 20 females). The mean age was 61.7 $\pm$ 7.0 (SD) years for male PD, 62.4 $\pm$ 9.2 years for female PD, 54.9 $\pm$ 9.7 years for male controls and 54.8 $\pm$ 8.1 years for female controls. PD subjects were pharmacologically treated and the voice recordings occurred at different moments of the day: patients could be on the effect on their treatment (ON-state) or not (OFF-state). The study was sponsored by INSERM,

and conducted according to Good Clinical Practice guidelines. All participants provided written informed consent and the research received approval from a local ethical committee and regulatory agencies.

B. Voice Recordings

The participants were recorded once in a consultation room at the hospital Pitié-Salpêtrière in Paris. Their voice was recorded with a professional head mounted omnidirectional condenser microphone (Beyerdynamics Opus 55 mk ii) placed approximately 10 cm from the mouth. This microphone was connected to a professional sound card (Scarlett 2i2, Focusrite). Speech was sampled at 96000 Hz with 24 bits resolution, with a spectrum of [50Hz, 20kHz]. There was a preamplification in the external sound card connected to the head mounted microphone.

C. Speech Tasks

The speech tasks were presented to the patients via a user interface on a computer programmed with Matlab. They had 18 short tasks to carry out. Each task lasted between 2s and 1min, leading to recordings of around 4min per subject. The tasks appeared in a random order. Some tasks started with an audio example. The speech tasks are presented in Table I. More details about our choice of tasks can be found in [25].

TABLE I
SPEECH TASKS WITH DESCRIPTION, OCCURENCE AND DURATION.

Task	Description	Occurrence	Duration[a]
aaa	Sustained vowel /a/ as long and steady as possible without breathing	2	20 s
DDK[b]	Repetition of syllables as fast and steadily as possible without breathing (/pa/, /pu/, /ku/, /pupa/, /paku/, /pataka/, /badaga/, /patiku/, /pabiku/, /padiku/)	all : 1 except /pataka/ : 2	2 min
reading	Reading of 2 short sentences (one question and one exclamative sentence) , 1 short text and 1 dialog.	1 each	1 min
free speech	Telling about one's day	1	1 min

[a] The duration is the total duration (all the occurrences cumulated for one type of task) and is an average among the participants

[b] DDK stands for diadochokinesia

III. METHODS: ACOUSTIC ANALYSIS AND CLASSIFICATION

A. Mel-frequency Cepstral Coefficients Extraction

MFCC are acoustic features introduced for the first time in 1980 by Davis and Mermelstein [26], for automatic speech recognition. Since then, they have been widely used in speech and speaker recognition, and more recently in the assessment of vocal pathologies. They characterize the spectrum envelop which directly depends on the shape of the vocal tract. We extracted MFCC with the Matlab Voicebox toolbox [27], as follows:

- *Framing*: Speech data are windowed into short frames. We chose 20ms frames, with an overlapping of 50% between two consecutive frames, and we applied Hamming windowing to avoid discontinuities between consecutive frames. This means a 10ms time shift between two consecutive frames.
- *Fast Fourier Transform (FFT)*: FFT, which is a fast way to compute Discrete Fourier Transform, is then calculated for each frame, resulting in one periodogram estimate of the power spectrum per frame.
- *Mel-spaced filterbank*: In order to get rid of useless information and get closer to the auditory perception, which perceives better the frequency variations in the low part of the spectrum, we applied Mel-spaced filter banks. The energy of the power spectrum is summed up inside portions of periodogram spectrum, filtered with overlapping triangular filters. These filters are sized and spaced along the mel-scale defined by:

$$Mel(f) = 1127 * ln(1 + f/700) \quad (1)$$

with f the frequency.

We used 34 filters for our implementation, so 34 filterbank energies characterize our power spectrum per frame.

- *Logarithm of filterbank energies*: As the human auditory perception of the intensity follows a logarithmic rule, the logarithm of the filterbank energies was taken.
- *Discrete Cosine Transform (DCT)*: DCT is applied on the log filterbank energies so that the main information concerning the envelop of the log filterbank energies lies in the first coefficients of the cepstrum computed with DCT. These coefficients are called MFCC, and we kept the first 12 of them.

To summarize, the first 12 MFCC are calculated from 20ms frames with 10ms time shifts.

B. Classification

The goal of our study is to be able to distinguish early PD patients from controls, knowing the gender of an unknown subject. We had speech data from early PD patients (males and females) and healthy controls (males and females) that we could use to build a statistical multidimensional GMM model per task for each of the 4 groups. Knowing the gender of a new subject we wanted to see with which accuracy we could predict if she/he belonged to the PD or control side. During the training phase, we built 12-dimensional GMM models (one per group and per task) which fit the MFFC distributions, as illustrated in Fig. 1. Means, SD and weights of the Gaussians (characterizing the GMM) were estimated via an Expectation-Maximization algorithm. We could choose the number of Gaussian functions used to fit the distribution of the MFCC. This number depended on the quantity of speech data we had per group (which depended on the number of subjects per group and the length of the task). Preliminary tests showed that 20 Gaussians seemed relevant for approximately 20 subjects per group for a 1min task (i.e. 20min of speech data).

To classify one test subject, we computed the log-likelihood of its MFCC values during one task against the PD and the control models (that corresponded to the same task and to his or her gender). The model against which the test subject

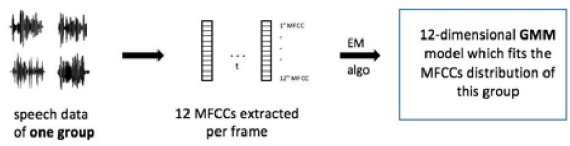

Fig. 1. Training phase: GMM model building for each group (male Parkinson, female Parkinson, male control and female control) and for each task. EM algo: Expectation-Maximization algorithm

had the highest log-likelihood would determine his/her class. More precisely, we computed one log-likelihood per frame of the data test, against the 2 models. Then we averaged the log-likelihoods on all the frames (Fig. 2.). This method guaranteed the likelihood to be independent of the frame number.

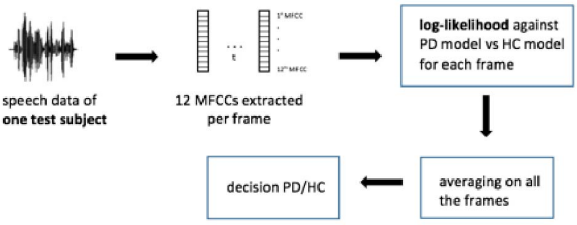

Fig. 2. Test phase: each test subject's data are tested against a PD model and a HC model for each task. PD: Parkinson's Disease, HC: Healthy control

C. Validation

As the number of subjects was not large enough, we could not split subjects into two groups: one for the training and test and one for the validation. Therefore, to still have a correct estimation of the real accuracy of our classifier, we used a leave one subject out (LOSO) cross validation. We used all the subjects, except one, for the training of our models and we used the remaining subject for the test. We repeated the procedure so that each subject was tested once. Then we computed the accuracy of the cross validation classification *(Acc_cv)* defined as the rate of well classified subjects.

The real accuracy (the one we would obtain on an infinite number of new test subjects) can be approximated by the cross validation accuracy we just computed. We estimated the density probability of the real accuracy by a Gaussian function with a mean equal to the cross validation accuracy, and a standard deviation defined by:

$$SD = \sqrt{Acc\_cv * (1 - Acc\_cv)/n} \quad (2)$$

with n being the number of subjects tested during the cross validation.

Actually this provides a good estimation of the real accuracy when there are enough tests (> 30) and when the tests are iid (independent and identically distributed) which is not completely the case in a cross validation (the tests are not completely independent). But for stable inducers we can use those formulas for LOSO cross validation [28]. In order to have an estimation of the true positives and the true negatives we also computed the sensitivity (rate of PD correctly classified) and the specificity (rate of controls correctly classified) of our classifier. The estimations of the real sensitivity and specificity follow the same rule as for the accuracy.

D. Statistical tests

In order to compare the performances obtained with the LOSO we conducted several statistical tests. As our number of subjects tested in each classification test was higher than 30 we used Welch's t-test to compare the accuracies of different populations (e.g. men vs women). We used a paired t-test to compare the accuracies for different tasks carried out by a same population. The p-value was computed each time and gave us information about the significance of the differences. We fixed the risk at 5%. That means we considered differences as significant when p-values were smaller than 0.05.

IV. Results

We performed the acoustic analyses and the LOSO cross validation on the 4 following tasks: a sustained phonation vowel task, a DDK task, a reading task and a free speech task. As mentioned previously, we tested men and women separately, because of the influence of the gender on the MFCC, and together (for the comparison). To avoid a gender bias, we removed (at random) 2 male PD and 6 female controls for the mixed-gender classification, so that the PD mixed-gender model and the control mixed-gender model had the same amount of males and females each. We tested our classifier with various numbers of Gaussians for our models and we had confirmation that 20 Gaussians was the best number for our tasks. The results from the LOSO cross validation calculated with 20 Gaussians for our models are presented in Table II.

Our classifier, based on a MFCC distribution modeling with GMM of healthy subjects and early PD patients, showed the best scores for the reading task (Acc=91.4% for men and 79.5% for women). The differences were significant when compared with the sustained vowel task for both men (p=0.0004) and women (p=0.03), and when compared with the free speech tasks for men (p=0.03).

As the sustained phonation task was shorter than the reading task we tested if its low classification performance could be explained by the short duration. Indeed, in general, the more speech data we have, the better we can train the GMM model and the higher is the performance of the classification. To check the influence of the duration of the task on the performance, we performed a classification on a 4s excerpt of the reading task (we just kept the 2 short sentences). After performing LOSO cross validation we obtained an accuracy of 88.6% for males and 74.4% for females. These scores are a little lower than the scores for the entire reading task, even though the difference is not significant (p=0.57 for men and p=0.42 for women). This slight difference is coherent with the positive role of the duration in the performance of the classifier. This shorter reading task still yields better results than the other tasks, and far better results than the sustained vowel task, confirming the relevance of a reading task for our classifier. Therefore we can conclude that the short duration of the sustained vowel task was not the main cause of its low score.

The comparison of the scores between males and females when they are confronted with same-sex groups indicates

globally higher scores for men. Nevertheless the differences are not significant (e.g. p=0.14 for the reading task). Scores for males alone are also better than those obtained with mixed-gender groups (Acc=75.8% during the reading task). This difference is significant (p=0.03). Thus it seems better for men to be classified separately, but it is not so obvious for women. In order to tackle this issue, we calculated the accuracy when classifying females between a mixed-gender PD group and a mixed-gender control group. The accuracy was 78.8% ($\pm$ 7.1%), which is in the same range as the one obtained with female alone groups (no significant difference: p=0.95 for the reading task). By contrast, the difference between the accuracies obtained for men classified with male models, and for men classified with mixed models, was significant (p=0.04 for the reading task). Therefore, treating separately men and women did not change the classification accuracy for women but significantly improved the men's one.

TABLE II
CLASSIFICATION PERFORMANCE OF EARLY PD AUTOMATIC DETECTION.

Tasks		PD men vs Control men	*SD*	PD women vs Control women	*SD*	PD vs Controls	*SD*
	Acc	60.0	8.3	59.0	7.9	48.5	6.2
aaa	Se	71.4	9.9	42.1	11.3	55.3	8.1
20s	Sp	42.9	13.2	75.0	9.7	39.3	9.2
	Acc	82.9	6.4	66.7	7.5	74.2	5.4
DDK	Se	81.0	8.6	57.9	11.3	71.1	7.4
2min	Sp	85.7	9.4	75.0	9.7	78.6	7.8
free	Acc	74.3	7.4	71.8	7.2	65.2	5.9
speech	Se	85.7	7.6	78.9	9.4	71.1	7.4
1min	Sp	57.1	13.2	65.0	10.7	57.1	9.4
	Acc	**91.4**	**4.7**	**79.5**	**6.5**	**75.8**	**5.3**
reading	**Se**	**95.2**	**4.6**	**78.9**	**9.4**	**84.2**	**5.9**
1min	**Sp**	**85.7**	**9.4**	**80.0**	**8.9**	**64.3**	**9.1**
reading	Acc	88.6	5.4	74.4	7.0	78.8	5.0
(extract)	Se	95.2	4.6	73.7	10.1	78.9	6.6
4s	Sp	78.6	11.0	75.0	9.7	78.6	7.8

Results are presented in %. Acc stands for accuracy, Se for Sensitivity, Sp for Specificity, SD for Standard Deviation, PD for Parkinson's Disease, aaa for sustained vowel /a/, DDK for diadochokinesia. Classification results are obtained with Leave one subject out cross validation.

V. Discussion

We adapted a method widely used in speaker recognition to early PD detection. This method consists in modeling MFCC distribution with GMM. First, a GMM was created for each group (PD and control). Then, the test subjects' MFCC were compared to these two models by means of log-likelihood functions, thus determining the classification between these two groups. GMM were here used as a classification component and not as a feature extraction step. This method doesn't appear to have been used in the past for PD detection. It showed results in line with the best performances found in the literature: around 90% for the detection of early stages of PD in men [6]. In this study [6] the performance obtained with just MFCC (but with another method) reaches 85% for early PD detection.

According to our study, the method we used seems to be specifically relevant for the reading task (Acc=91% for men), and especially for the reading of short questions or exclamation sentences (Acc=88% for only 4 sec of speech data per men). The fact that the reading task yielded a better performance than the free speech task with an equal duration may be explained by the fact that in the reading task, the variability depends only on the speakers' voice and on the effect of PD on their voice, whereas in the free speech task the variability depends also on the linguistic content (the sentences and words chosen). The sustained vowel task seemed particularly inappropriate for our analysis method (Acc=60% for men, that is just a little better than random would do).

Our study revealed that splitting males and females increased the classification performance. Actually it increased particularly male classification performances. The lower performance obtained for women as compared to men, may be explained by the fact that, in the cepstral domain analysis, female voices seem to have wider distributions and thus are more difficult to classify [23]. The same effect can be seen for the detection of other voice disorders, such as laryngeal disorders [23].

We may have obtained even better performances if the PD subjects were off medication because the PD medications improve vocal perturbations [29].

In the future we will record more subjects (PD and controls) and add some new vocal tasks. We will add delta MFCC, because studies showed delta MFCC could carry additional information for the assessment of parkinsonian voices' perturbations [30]. Finally we will combine the classification method presented in this paper, based on GMM and short term MFCC features, with more classical methods using global features that are discriminant in the detection of early stages of PD (cf annex 2 of [25]).

VI. Conclusion

An automatic and simple methodology, based on the one classically used for speaker recognition, was presented to detect early stages of Parkinson's disease. We performed an automatic analysis of the recordings of early PD patients and healthy controls for 4 types of tasks: sustained vowel, DDK, free speech and reading. The analysis methodology consisted in extracting 12 MFCC every 10ms for each task, and fitting their distribution with 2 multi-dimensional GMM: one for the PD model and one for the control model. The log-likelihood of test subjects' MFCC values against those two models allowed a 2-class classification between PD group and control group. The results obtained with a LOSO cross validation seem to indicate that the reading task is particularly appropriate for this methodology. The accuracy obtained for this task (91% for men) is in line with the best performance that we can find in the literature concerning early PD detection. Combining this analysis method with other more classical ones used in PD detection may improve classification performance for women and for the other vocal tasks.

VII. Acknowledgment

L. Jeancolas was supported by a grant of Institut Mines-Télécom, Fondation Télécom and Institut Carnot Télécom & Société Numérique through "Futur & Ruptures" program. The ICEBERG study was partly funded by the program "Investissements d'Avenir" ANR-10-IAIHU-06, and Fondation EDF. The authors would like to thank ICM and CIC Neurosciences, unité des Pathologies du sommeil and CENIR teams, especially E. Benchetrit and I. Arnulf for their helpful cooperation.

REFERENCES

[1] L. M. De Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474442206704719

[2] J. Rusz, R. Čmejla, H. Růžičková, and E. Růžička, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, p. 350, 2011. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/129/1/10.1121/1.3514381

[3] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and Cognition*, vol. 56, no. 1, pp. 24–29, Oct. 2004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0278262604001393

[4] R. B. Postuma, A. E. Lang, J. F. Gagnon, A. Pelletier, and J. Y. Montplaisir, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012. [Online]. Available: http://brain.oxfordjournals.org/content/135/6/1860

[5] J. Rusz, J. Hlavnička, T. Tykalová, J. Bušková, O. Ulmanová, E. Růžička, and K. Šonka, "Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder," *Sleep Medicine*, Sep. 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1389945715009296

[6] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016. [Online]. Available: http://asa.scitation.org/doi/abs/10.1121/1.4939739

[7] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[8] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, vol. 1, 2002, pp. 182–183 vol.1.

[9] J. Godino-Llorente and P. Gómez-Vilda, "Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, Feb. 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1262116/

[10] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–873.

[11] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6126094

[12] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. V. Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of parkinson's disease from words uttered in three different languages." in *INTERSPEECH*, 2014, pp. 1573–1577.

[13] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolaños, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, and E. Nöth, "Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.

[14] J. R. Orozco-Arroyave, J. C. Vàsquez-Correa, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *ICASSP 2016*. IEEE, Mar. 2016, pp. 6490–6494. [Online]. Available: http://ieeexplore.ieee.org/document/7472927/

[15] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic Detection of Parkinson's Disease Based on Modulated Vowels," in *Interspeech 2016*, Sep. 2016, pp. 1190–1194.

[16] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating Between Patients With Parkinson's and Neurological Diseases Using Cepstral Analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 10, pp. 1100–1108, Oct. 2016.

[17] A. Jafari, "Classification of Parkinson's Disease Patients using Nonlinear Phonetic Features and Mel-Frequency Cepstral Analysis," *Biomedical Engineering: Applications, Basis and Communications*, vol. 25, no. 04, p. 1350001, Aug. 2013. [Online]. Available: http://www.worldscientific.com/doi/abs/10.4015/S1016237213500014

[18] A. Benba, A. Jilbab, and A. Hammouch, "Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 449–456, Sep. 2016. [Online]. Available: http://link.springer.com/10.1007/s10772-016-9338-4

[19] T. Kapoor and R. K. Sharma, "Parkinson's disease diagnosis using Mel-frequency cepstral coefficients and vector quantization," *International Journal of Computer Applications*, vol. 14, no. 3, pp. 43–46, 2011.

[20] A. Benba, A. Jilbab, and A. Hammouch, "Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ," in *The 2014 international conference on circuits, systems and signal processing*, 2014, pp. 23–25. [Online]. Available: http://www.inase.org/library/2014/russia/bypaper/CCCS/CCCS-15.pdf

[21] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. Upper Saddle River, NJ: Prentice Hall, Nov. 2001.

[22] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." in *Interspeech*, 2013, pp. 1149–1153. [Online]. Available: http://ai2-s2-pdfs.s3.amazonaws.com/ecfe/10780a4e45031024860068d8cc98b78abb44.pdf

[23] R. Fraile, N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 146–152, Jul. 2009.

[24] A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo, and A. J. Lees, "The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service," *Brain*, vol. 125, no. 4, pp. 861–870, Apr. 2002. [Online]. Available: http://brain.oxfordjournals.org/content/125/4/861

[25] L. Jeancolas, D. Petrovska-Delacrétaz, S. Lehéricy, H. Benali, and B.-E. Benkelfat, "L'analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson : état de l'art," in *CORESA 2016 : 18e Edition COmpressions et REpresentation des Signaux Audiovisuels*. Nancy: CNRS, May 2016, pp. 113–121.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[27] "VOICEBOX." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[28] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14. Stanford, CA, 1995, pp. 1137–1145. [Online]. Available: https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf

[29] J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, V. Majerová, J. Picmausová, J. Roth, and E. Růžička, "Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy," *Journal of Neural Transmission*, vol. 120, no. 2, pp. 319–329, Feb. 2013. [Online]. Available: http://link.springer.com/10.1007/s00702-012-0853-4

[30] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011. [Online]. Available: http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2010.0456