INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic detection of Parkinson's disease based on modulated vowels

Daria Hemmerling[1], *Juan Rafael Orozco-Arroyave*[2,3], *Andrzej Skalski*[1], *Janusz Gajda*[1], *Elmar Nöth*[3]

[1] AGH University of Science and Technology, Department of Measurement and Electronics,
Al. Mickiewicza 30, 30-059 Krakow, Poland
[2]Department of Electronics and Telecommunications Engineering, Universidad de Antioquia UdeA,
Calle 70 No. 52-21 Medellín, Colombia
[3]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

hemmer@agh.edu.pl, rafael.orozco@udea.edu.co

Abstract

In this paper we present a novel approach of automatic detection of phonatory and articulatory impairments caused by Parkinson's disease (PD). Modulated (varying between low and high pitch) and sustained vowels are considered and analysed. The fundamental frequency of the phonations and its range are computed using the Hilbert-Huang transformation. Additionally, a set with "standard" measures are calculated to model phonatory and articulatory deficits exhibited by Parkinson's patients. Kernel Principal Component Analysis was also applied in order to reduce the dimensionality of the representation space. The automatic discrimination between speakers with PD and healthy controls (HC) is performed using decision trees. According to the results, modulated vowels are suitable to evaluate phonatory and articulatory deficits observed in PD speech.

Index Terms: Parkinson's disease, pitch modulation, decision trees, kernel Principal Component analysis, phonatory, articulatory.

1. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder in the world affecting about 2% of people older than 65 years [1]. Patients with PD develop several motor impairments such as bradykinesia, rigidity, postural instability, and resting tremor. Non-motor deficits developed by PD patients include sleep disorders and problems with cognition and emotion [2]. The majority of PD patients develop several speech impairments with symptoms including monoloudness, breathy, hoarse and rough voice, inappropriate pauses, misarticulation, and trembling voice [3]. Different studies have reported instability in the phonation of PD speakers and articulatory deficits that affect the speech production and its intelligibility [4, 5, 6]. Phonatory and articulatory deficits can be studied with sustained phonations or continuous speech signals. Most of the studies about Parkinson's speech consider sustained phonations because this speech task is easy to be reproduced by elderly patients and because it provides information about the phonatory (vibration of the vocal folds) and articulatory (resonances in the vocal tract) processes of speech production. The abnormalities in phonation of PD patients have been widely studied considering the fundamental frequency, its variability, and also nonlinear dynamics techniques have been introduced to model such impairments [7]. Regarding the analysis of articulatory abnormalities, they have been also studied mainly considering sustained phonations of the vowels /a/, /i/, and /u/. In [8] and [9] the authors introduce the Formant Centralization Ratio (FCR) as a new measure to assess the articulatory capability of PD speakers. FCR is defined as $\text{FCR} = \frac{F_1/i/+F_1/u/+F_2/u/+F_2/a/}{F_2/i/+F_1/a/}$, where $F_1/a/$, $F_1/i/$, and $F_1/u/$ are the frequency of the first formant of the vowels /a/, /i/, and /u/. Similarly, $F_2/a/$, $F_2/i/$, and $F_2/u/$ are the frequency of the second formant of the vowels /a/, /i/, and /u/, respectively. According to the results, FCR is suitable to robustly differentiate dysarthric from healthy speech. In [10] the triangular Vowel Space Area (tVSA) is evaluated to assess articulation in PD speakers. tVSA is defined as the area of the triangle that is formed when the first two formants extracted from the corner vowels /a/, /i/, and /u/ are plotted in the (F_1, F_2) plane. In paper [11] the authors present phonation, articulation and prosody analyse based on: sustained vowels, vowels uttered with changing the tone of each vowel from low to high, different words, phonemes, and different speech tasks. The authors used PC-GITA database, the same as was used in following paper. The highest classification accuracy to detect Parkinson's disease was 91,3% obtained with the vowel /a/ modelled with periodicity and stability measures. Recently, in [6] the authors present a study where classical and nonlinear dynamic features are applied to analyze Parkinson's speech. The highest classification accuracy being at the level of 91% was achieved for vowel /a/ using stability and periodicity features. According to the results, phonatory-based features are more suitable to study problems in PD speech when sustained vowels are analyzed; however, note that studies with modulated vowels, i.e., sustained vowels uttered changing their tone from low to high, have not been addressed to detect PD so far. The study of this kind of speech tasks is relevant because they can reflect difficulties to modulate the tone of the phonation and also show deficits in moving the tongue to the correct position to produce modulated vowels.

In this paper we consider recordings of the Spanish vowel /a/ uttered in a sustained manner and with modulated pitch. The voice signals are modeled with several features to describe phonatory and articulatory phenomena which are impaired in Parkinson's speech. Kernel Principal Component Analysis (kPCA) is used to eliminate redundant information and to remove correlation among features. The automatic discrimination of PD and HC speakers is performed with decision trees. The results show that articulatory-based features complement the information provided by the phonatory-based measures, which improves the classification accuracy. The rest of the paper is as follows: section 2 include the voice database

http://dx.doi.org/10.21437/Interspeech.2016-1062

description, section 3 showcases the methodologies used in the examination including phonatory modeling, articulatory modeling, feature space transformation and classification. Section 4 presents obtained results. Section 5 contains the conclusion of all the work featured in this paper.

2. Voice Database

Recordings of the PC-GITA database [11] are considered. A total of 100 speakers (50 with PD and 50 HC) are included. All the participants are Colombian Spanish native speakers. The age of the men with PD ranges from 33 to 77 years old (mean 62.2 ± 11.2), the age of the women with PD ranges from 44 to 75 years old (mean 60.1 ± 7.8). For the case of healthy speakers, the age of the men ranges from 31 to 86 (mean 61.2 ± 11.3) and the age of the women ranges from 43 to 76 years old (mean 60.7 ± 7.7). The sampling frequency of the recordings was 44.1 kHz with 16-bit resolution. The speakers were asked to produce two different speech tasks, (1) sustained phonations of the vowel /a/ at a constant tone, and (2) sustained vowels with a modulated tone, i.e., varying from low to high. All of the patients were recorded in ON-state, i.e., no more than three hours after their morning medication. The neurological state of the patients was assessed by a neurologist expert according to the MDS-UPDRS-III (mean 38.5 ± 19.1) scale [12].

3. Methodology

3.1. Phonatory modeling

Several measures typically used for modeling phonatory and articulatory deficits are used. The phonatory model includes mean value, standard deviation, kurtosis, and skewness of the fundamental frequency (F_0) and the difference between its maximum and minimum per phonation. The mean values of jitter, shimmer, the curvature of the pitch-contour, 10 mel-frequency cepstral coefficients (MFCCs), and energy contour are also considered. Further details of the algorithms applied to compute these features can be found in [13, 14]. Additionally, the instantaneous energy and its range are computed using the Hilbert-Huang transformation (HHT) [15]. The HHT is an extension of Empirical Mode Decomposition (EMD) algorithm by applying the Hilbert transform. It allows the determination of instantaneous frequencies and amplitude components of a decomposed signal. The use of Hilbert spectral analysis facilitates the isolation of signal's subsequent components to determine which of them are more prominent and “dominate” the frequency of the signal. In the HHT procedure no assumption about the stationarity of the signal is made, thus this transformation can provide a higher resolution than short-time Fourier transform (STFT) for analyses in the time-frequency domain. Figure 1 shows the difference between the spectrogram obtained with the HHT transform (part A of the figure) and with the STFT transform (part B of the figure). Note that HHT provides more detailed information of the signal in the low frequency bands (between 100 Hz and 300 Hz).

The HHT transformation was introduced in [15]. The authors present an adaptive technique to represent signals as a sum of simpler components in the time domain. Such components are obtained by using the EMD which allows their separation in time and frequency. The first step of the HHT computation consists of a low-pass filtering with a sharp cut-off frequency of 800 Hz in order to shorten the frequency-band to those values where pitch values exist [16] . Then a set of decomposed functions which are called intrinsic mode functions (IMF) is formed. To calculate IMF the following conditions must be fulfilled: the number of extrema and the number of zero-crossing must be equal or differ at most by one and an average value of the envelope interpolating local maxima and the envelope interpolating local minima is equal to zero. To obtain the IMF functions the procedure known as *sifting* is implemented as follows. Firstly all the extrema (local maxima and minima) of the signal $x(t)$ are identified. Secondly, the mean value $m(t)$, the upper- and lower-envelopes of the signals $x(t)$ are obtained. Thirdly the difference $d(t) = x(t) - m(t)$ is calculated to extract the detailed signal. In fact, $d(t)$ rarely satisfies the two conditions mentioned before for the IMF functions. Therefore, the *sifting* procedure must be repeated several times, with the “difference” $d(t)$ taking the place of $x(t)$. To calculate the next IMF, the entire process is applied to the residual signal $r_1(t) = x(t) - d(t)$. The residual is iterated until two conditions are satisfied: (1) the number of extrema in the residual is smaller than 2 and (2) the maximum number of iterations is reached. The role of EMD is to decompose an arbitrary and time-varying signal into intrinsic mode functions that are modulated in amplitude and frequency. The IMFs represent actual fluctuations of the signal (including both variables: amplitude and frequency as functions of time). The sum of IMFs gives the original signal. The Hilbert transformation is applied upon the IMFs and the instantaneous frequencies and amplitudes are found. Figure 2 shows the waveforms of acoustic signals and their first 4 IMFs (Part (A) indicates a healthy woman and part (B) a woman suffering from PD). Note that the IMF3 and IMF4 signals are different for the healthy and pathological speaker, i.e., the signals obtained from the patient are more choppy than those exhibited by the healthy speaker. This behavior indicates more complexity in the dynamical structure of Parkinson's voice, which was previously

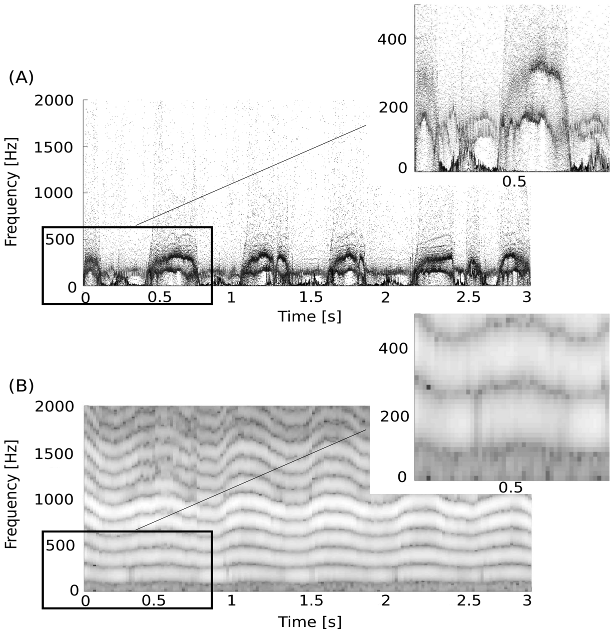

Figure 1: *Spectrograms of a modulated vowel /a/. Part (A) corresponds to the HHT and part (B) corresponds to the STFT. Patient information: female, 55 years old, MDS-UPDRS-III 43, and recorded after 12 years of PD diagnosis.*

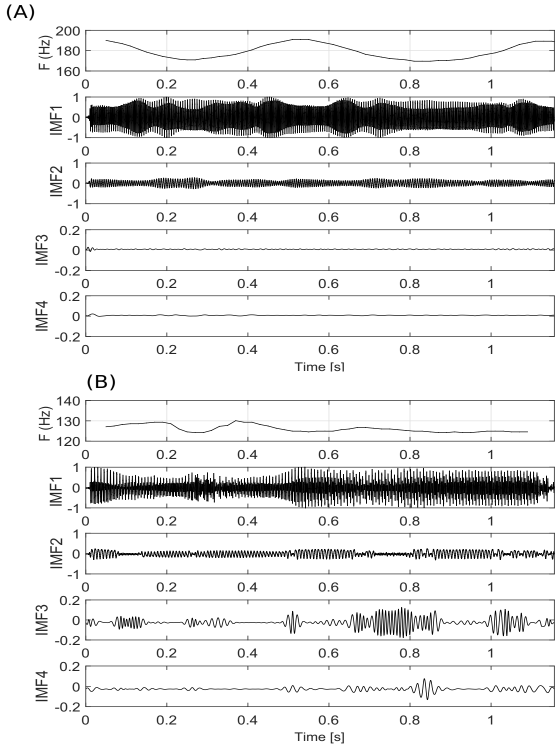

Figure 2: *Pitch changes of acoustic signal and first 4 IMFs for A) healthy woman 63 years old, B) woman with Parkinson Disease 57 years old, 41 UPDRS, 3 H&Y, 37 years after diagnosis*

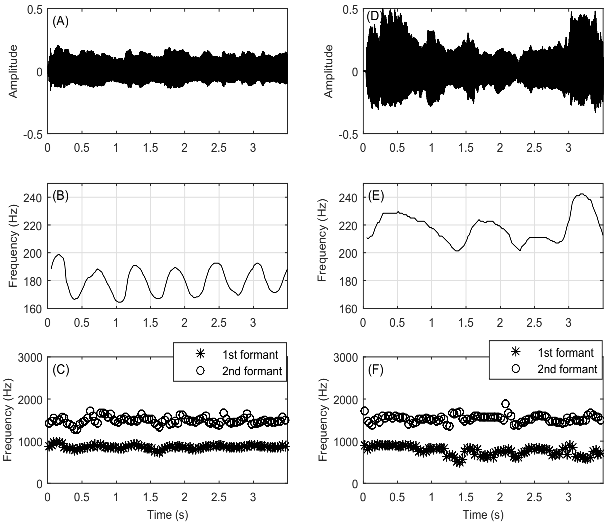

Figure 3: *A) Waveform of 'a' vowel signal, healthy woman (HC), age 61, B) pitch modulation HC, C) first two formant's frequencies representation, HC, D) waveform of 'a' vowel signal, woman with Parkinson's disease (PD), age 60, UPDRS=29, H&Y=2, 7 years after diagnosis, E) pitch modulation PD, F) first two formant's frequencies representation, PD*

observed in [17] as an index of phonatory instability and voice tremor of PD speakers.

The instantaneous frequency of a signal can be obtained from analytic function defined as $z_i(t) = c_i(t) + jd_i(t) = a_i(t)e^{j\delta_i(t)}$, where $a_i(t)$ indicates the instantaneous amplitude, and $\delta_i(t)$ represents the instantaneous phase of the IMF $c_i(t)$. The instantaneous frequency $\omega_i(t)$ of $c_i(t)$ is calculated as $\omega_i(t) = d\delta_i(t)/dt$, thus the acoustic signal $x(t)$ can be represented as:

$$x(t) = \mathbb{R}\left\{\sum_{i=1}^{n} a_i(t)e^{j\omega_i(t)}\right\} \qquad (1)$$

where $\mathbb{R}\{\cdot\}$ indicates the real part of the argument. Instantaneous frequencies outside the range between 60 Hz and 500 Hz are set to zero. Similarly, values with variation greater than 100 Hz within 5 ms are also set to zero [18]. The IMF with the largest amplitude is chosen for the computation of the instantaneous frequency. The difference between the maximum and the minimum values of the IMFs amplitude is also computed.

3.2. Articulatory modeling

Regarding the articulatory modeling we computed the mean values of the first two formants (F_1 and F_2) which provide information about possible instabilities or abnormalities in the shape of the vocal tract and also about the vertical and horizontal position of the tongue [19, 20]. As during the production of speech Parkinson's patients exhibit articulatory movements restricted in range and energy due to hypokinetic dysarthria, we hypothesize that such abnormalities can be observed more clearly during the production of modulated vowels, thus the frequency formants could provide information regarding instabilities and abnormal positions of the vocal tract during such phonations.

Figure 3 shows different waveforms obtained from one healthy speaker (left side: figures A, B, and C) and one PD speaker (right side: figures D, E, and F). Parts A and D indicate the acoustic signal of the modulated vowel, parts B and E show the pitch-contour, and parts C and F show the contour of the first two formants. From the plots of the pitch-contour we can observe that it was difficult for PD patients to perform the modulation of the vowel. Additionally, the instability in the movement of the tongue while producing modulated vowels can be observed in the contours of the first two formants. Note that there is a couple of hops or discontinuities specially in the first formant obtained from the PD patient (part F).

3.3. Feature space transformation and classification

Kernel Principal Component Analysis (KPCA) is used here to reduce the dimensionality of the feature space without any assumption regarding possible linear relationships among the computed features [21, 22]. A Gaussian kernel is used here and its parameter σ is optimized such that it should be smaller than inter-class distances and larger than inner-class distances, thus it is calculated as follows [22]:

$$\sigma = \xi \times mean(l_i^{\mathrm{NN}}) \qquad (2)$$

where ξ was optimized during the training process to achieve the best data separation, l_i^{NN} is the distance from analyzed data point to its nearest neighbor. The computations were done for female and male recordings separately, thus we obtained different σ values on each case.

A decision tree was used to perform the automatic discrimination between PD and HC speakers. Decision trees are used

Table 1: *Results with phonatory features*

	sustained vowel			
	original vector		kPCA	
	female	male	female	male
Accuracy	72±4%	76±5%	90±4%	80±6%
# of features	23	23	17	14
	modulated vowel			
Accuracy	64±4%	78±5%	76±4%	82±4%
# of features	23	23	15	11

Table 2: *Results with articulatory features (F_1 and F_2)*

	sustained vowel			
	original vector		kPCA	
	female	male	female	male
Accuracy	57±4%	60±3%	58±4%	60±6%
	modulated vowel			
Accuracy	60±4%	64±4%	64±2%	64±6%

Table 3: *Results with articulatory and phonatory features*

	sustained vowel			
	original vector		kPCA	
	female	male	female	male
Accuracy	72±4%	76±4%	90±4%	82±4%
# of features	25	25	12	24
	modulated vowel			
Accuracy	68±3%	78±4%	76±4%	84±5%
# of features	25	25	17	25

to predict that each observation belongs to the most commonly occurring class of training observations in the region it belongs to [23]. The result of classification with decision trees is binary, i.e., 0 for healthy and 1 for PD. A 10-fold cross-validation (CV) strategy is followed for the training process, i.e., the data is split into 10 folds randomly chosen, 9 of them are used for train and 1 is for test. The procedure is repeated 10 times. Each subject is in a different test fold, and the same subject never is in both test and train groups. It then examines the predictive accuracy of each new tree on the data not included in training that tree. Therefore, validation set is used to avoid overfitting.

The optimal tree size is chosen by selecting the smallest error rate obtained in the test set.

4. Experiments and results

Phonatory and articulatory analyses are performed upon sustained and modulated vowels. Sustained vowels are included for the sake of comparisons with respect to the modulated vowels. The results of the 2-class problem of detecting whether the patient is suffering from PD or not are summarized in Tables 1, 2, and 3. Table 1 includes only results obtained with phonatory features, Table 2 indicates the results obtained with only articulatory features, i.e., F_1 and F_2. Table 3 shows the results obtained when phonatory and articulatory features are combined.

The results in Tables 1 and 2 indicate that at least with the feature sets considered here, the phonatory modeling is more suitable to assess sustained and modulated vowels than the articulatory modeling. However, Table 3 shows that when considering recordings of the male speakers and the kPCA procedure, the results can improve. This result indicates that, to some extent, the phonatory and articulatory features are complementary and it is worth to consider the combination of these two feature sets.

5. Conclusions

In this paper we have analyzed recordings of 50 PD patients and 50 healthy control speakers. The participants performed the sustained phonation of the Spanish vowel /a/ with a constant tone and also with a modulated tone, i.e., varying from low to high. Phonatory and articulatory features are extracted from the recording. Phonatory modeling includes several measures such as jitter, shimmer, F_0, energy, and MFCCs. Additionally, the instantaneous frequency and its range are estimated with the HHT transformation. Regarding the articulatory measures, the first two formants are computed in order to model deficits of PD patients to move and hold the tongue and jaw in their correct position while pronouncing sustained and modulated vowels. Two versions of the feature sets are considered: one with all of the computed features and another one with the result of applying the kPCA transformation. The automatic discrimination of PD people and HC speakers is performed with a decision tree. The results obtained with the phonatory features are better than those obtained with the articulatory features. This result indicates that phonatory modeling is more suitable to assess sustained phonations of the vowel /a/ with constant and modulated pitch. The results obtained with the articulatory features show that the considered feature set (only with the first two formants) is not suitable to model articulatory impairments in PD speech. The best classification results obtained with male and female considering sustained vowels are 82% and 90%, respectively. Regarding the results with the modulated vowels, an accuracy of 76% was obtained in female and 84% in male.

This is a preliminary study that considers a new speech task (with modulated vowels) to assess phonatory and articulatory features in the speech of PD patients. The modulated vowels enable an extension of the standard approach of the acoustic analysis, adding articulatory and phonatory information against differences in age and gender. Further research is required to include more features in order to take more advantage of the proposed speech task, which could be useful to evaluate problems of PD patients to keep their tongue and jaw in the correct position while producing sustained and modulated phonations. Other papers ([6] and [11]) show that adding non linear features and other measurements such as shimmer and harmonics-to-noise ratio might increase the classification accuracy. Also exceeding the analysis to word, sentences, text and monologue might bring useful information to the classification process. For the future work we will extract more features from the HHT transformation, i.e., from the IMFs decompositions. According to the preliminary observations, these representation seems to be promising to model several phenomena in the low frequency range of the speech spectrum.

6. Acknowledgements

This work was funded by the Ministry of Science and Higher Education in Poland under the Diamond Grant program, Decision number 0136/DIA/2013/42 (AGH 68.68.120.364). This study was partially funded by COLCIENCIAS project # 111556933858.

7. References

[1] M. de Rijk, “Prevalence of parkinson’s disease in europe: A collaborative study of population-based cohorts,” *Neurology*, vol. 54, pp. 21–23, 2000.

[2] J. Logemann, H. Fisher, B. Boshes, and E. Blonsky, “Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients,” *Journal of Speech and Hearing Disorders*, vol. 43, pp. 47–57, 1978.

[3] L. O. Ramig, C. Fox, and S. Sapir, “Speech treatment for parkinsons disease,” *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[4] S. Skodda, W. Grönheit, and U. Schlegel, “Impairment of vowel articulation as a possible marker of disease progression in Parkinson’s disease,” *PloS one*, vol. 7, no. 2, 2012.

[5] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, “Imprecise vowel articulation as a potential early marker of Parkinson’s disease: effect of speaking task,” *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[6] J. R. Orozco-Arroyave, E. A. Belalcázar-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*, no. 6, pp. 1820–1828, 2015.

[7] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, “Suitability of dysphonia measurements for telemonitoring of parkinson’s disease,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[8] S. Sapir, L. O. Raming, J. L. Spielman, and C. Fox, “Formant centralization ratio (FCR): a proposal for a new acoustic measure of dysarthric speech,” *Journal of Speech Language and Hearing Research*, vol. 53, no. 1, pp. 1–20, 2010.

[9] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, “Acoustic metrics of vowel articulation in parkinson’s disease: vowel space area (vsa) vs. vowel articulation index (vai),” in *MAVEBA*, 2011, pp. 173–175.

[10] S. Skodda, W. Visser, and U. Schlegel, “Vowel articulation in parkinson’s diease,” *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011, Erratum in Journal of Voice. 2012 Mar;25(2):267-8.

[11] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, “New spanish speech corpus database for the analysis of people suffering from parkinson’s disease.” in *LREC*, 2014, pp. 342–347.

[12] C. G. Goetz and et al., “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results,” *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[13] D. Panek, A. Skalski, and J. Gajda, “Quantification of linear and non-linear acoustic analysis applied to voice pathology detection,” in *Information Technologies in Biomedicine, Volume 4*. Springer, 2014, pp. 355–364.

[14] D. Panek, A. Skalski, J. Gajda, and R. Tadeusiewicz, “Acoustic analysis assessment in speech pathology detection,” *International Journal of Applied Mathematics and Computer Science*, vol. 25, no. 3, pp. 631–643, 2015.

[15] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, “The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis,” in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971. The Royal Society, 1998, pp. 903–995.

[16] P. Veprek and M. S. Scordilis, “Analysis, enhancement and evaluation of five pitch determination techniques,” *Speech Communication*, vol. 37, no. 3, pp. 249–270, 2002.

[17] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, “Phonatory Impairment in Parkinson’s Disease: Evidence from Nonlinear Dynamic Analysis and Perturbation Analysis,” *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.

[18] H. Huang and J. Pan, “Speech pitch determination based on hilbert-huang transform,” *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006.

[19] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter, 1971, vol. 2.

[20] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, “Generating vocal tract shapes from formant frequencies,” *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.

[21] B. Schölkopf, A. Smola, and K.-R. Müller, “Kernel principal component analysis,” in *Artificial Neural NetworksICANN’97*. Springer, 1997, pp. 583–588.

[22] Q. Wang, “Kernel principal component analysis and its applications in face recognition and active shape models,” *arXiv preprint arXiv:1207.3538*, 2012.

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.