Automatic Evaluation of Parkinson's Speech - Acoustic, Prosodic and Voice Related Cues

Tobias Bocklet[1], *Stefan Steidl*[1], *Elmar Nöth*[1,2], *Sabine Skodda*[3]

[1]Chair of Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
[2]ECE, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]Knappschaftskrankenhaus Bochum, Bochum, Germany

tb@speech.cs.fau.de

Abstract

Articulation and phonation is affected in 70 % to 90 % of patients with Parkinson's disease (PD). This study focuses on the question whether speech carries information about 1. PD being present at a speaker or not, and 2. estimating the severity of PD (if present). We first perform classification experiments focusing on the automatic detection of PD as a 2-class problem (PD vs. healthy speakers). The detection of severity is described as a 3-class task based on the Unified Parkinson's Disease Rating Scale (UPDRS) ratings. We employ acoustic, prosodic and glottal features on different kinds of speech tests: various syllable repetition tasks, read sentences and texts, and monologues. Classification is performed in either case by SVMs. We report recognition results of 81.9 % when trying to differentiate between normally speaking persons and speakers with PD. With system fusion we achieved a recognition results of 59.1 % on the task of UPDRS classification.

Index Terms: Parkinson's Disease, pathologic speech, speech analysis

1. Introduction

Parkinson's disease (PD) is a degenerative disorder of the central nervous system. It results from the death of dopamine-containing cells in the substantia nigra, a region of the midbrain and is the second most common neurodegenerative disorder after Alzheimer's disease [1]. PD accounts for a variety of motor (shaking, rigidity, movement difficulties, and communication) and non-motor deficits (effects on the sensory system, sleep, and emotion) with speech being affected in between 70 % and 90 % of all PD patients [2]. Medical treatment alleviates certain symptoms, but there is no causal cure now available, and early diagnosis is critical for maximizing the effect of treatment and improving the quality of the patient's life [3].

Several speaking tasks have been developed for the evaluation of PD speech and voice. The most traditional of them are sustained phonation, rapid syllable repetition, variable reading of short sentences, longer passages and freely spoken spontaneous speech [4].

In a previous work [5] we focused on an automatic detection of PD speakers in early stages based on a small dataset of Czech speakers [6]. We used systems based on different levels of voice and speech (phonation, articulation and prosody) in order to evaluate which speech levels and speech tests are the most discriminative ones for an automatic detection (classification task with two classes) of PD speakers in early stages.

In this work we made use of a larger German dataset of 88 speakers with PD and 88 control speakers performing various PD-related speech tests. Additionally, the dataset contains labels of the *Unified Parkinson's Disease Rating Scale III* (UPDRS-III). UPDRS is a long-term questionnaire with 14 questions that evaluates Motor Symptoms of PD patients. The results of the questionnaire results in an integer scale ranging from 0 (no impairment) to 56 (high impairment). One of the questions focuses on the patient's speech, the UPDRS-III in general shows weak correlations with basic (prosodic) features derived from speech [7]. In the long term, this research should lead to an easy-to-use screening (does the person develeop PD) and surveillance (did the disease of a PD person get worse) system. Thus, we focused on a detection of PD speakers with various systems and applied a system- and test-based fusion in order to improve these 2-class results as well as on a classification of the UPDRS scale, treating the UPDRS scale as three classes.

Four differently motivated systems are used in this work: Phonation is modeled by a glottal excitation system based on two-mass vocal fold modeling, articulation is modeled by spectral features followed by statistical modeling, and the prosody of a speaker is evaluated by a language-independent prosodic analysis based on voiced/unvoiced (VUV) decision [8]. Additionally, openSMILE, an all-purpose system containing features of all three levels, is used.

The outline of this paper is as follows: The data and the different speech tasks are described in Sec. 2. The different modeling approaches are presented in Sec. 3. Classification and experimental results are presented in Sec. 4, followed by concluding remarks in Sec. 5. The paper ends with a short summary (Sec. 6).

2. Data

2.1. Patients

176 German native speakers participated in this study. 88 speakers (44m, 44w) were diagnosed with PD and received medical treatment. PD symptoms being present since 6.6 years (±5.8). The patients' mean age is 66.6 years (±9.0). 88 healthy speakers (45m, 43w) with no history of neurological or communication disorders act as control group. Their mean age was 58.1 years (±14.2). Age distribution showed no significant differences between both groups (see Fig. 1).

The speech data was recorded in the Knappschaftskrankenhaus in Bochum, Germany, in a quiet room with a low ambient noise level using an external condenser microphone. The voice signals were sampled at 44.1 kHz, with 16-bit resolution. The severity of Parkinson was evaluated regarding the the *Unified*

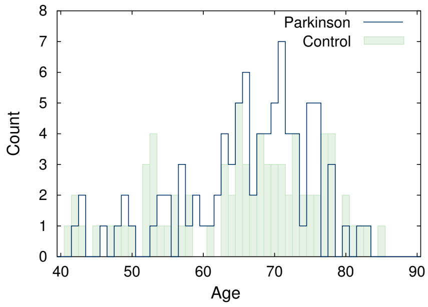

Figure 1: Distribution of ages between PD speakers (black) and control speakers (green)

Parkinson's Disease Rating Scale III (UPDRS-III). UPDRS-III is a long-term questionnaire with 14 questions that evaluates Motor Symptoms of PD patients. The results of the questionnaire results in an integer scale ranging from 0 (no impairment) to 56 (high impairment). One of the questions focuses on the patient's speech, The distribution among the PD speakers is shown in Fig. 2. In order to characterize the detection of UPDRS scores as classification problem, we assigned the speakers to three different UPDRS classes in order to achieve a dataset balanced regarding the number of speakers. The classes and number of according speakers are

- UPDRS score 0-15: 27 speakers
- UPDRS score 16-25: 32 speakers
- UPDRS score 26-55: 29 speakers

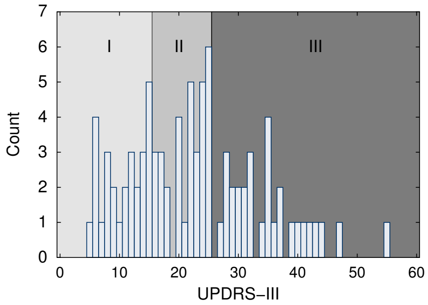

Figure 2: Distribution of UPDRS scores among the 88 PD speakers. The three different classes are shaded in gray.

2.2. Speech Tasks

Similar to the Czech dataset used in our previous work [5, 6], also the German dataset contains speech of different tasks. We selected a subset of these tasks in this work. These are:

- T01: spontaneous speech
- T02: read text (phonetically rich)
- T03: reading of Question-Answer-Pairs with stress on certain words
- T04: read sentences
- T05: read words
- T11: sustained vowel
- T44: repetition of single syllable (pa)
- T77: repetition of 3 syllables (pa-ta-ka)
- T88: repetition of syllable sequence (pa-pe pa-pi pa-po pa-pu)

3. Features

3.1. Acoustic Modeling

Gaussian Mixture Models (GMMs) model acoustic features, namely Mel Frequency Cepstrum Coefficients (MFCCs) in a statistical way. For acoustic feature extraction a Hamming window with a size of 25 ms and a time shift of 10 ms is applied to the speech signal. Afterwards the Mel-spectrum with 26 triangular filters is calculated and processed by Discrete Cosine Transform (DCT). We take the first 13 Mel-frequency Cepstral coefficients including C_0. Cepstral mean subtraction (CMS) is applied and first- and second order derivatives of these features are calculated over a context of 5 and 9 consecutive frames. In the end a 39-dimensional feature vector is created. This feature vector is then modeled by GMMs. For each speaker and speech task, one GMM is created by GMM-UBM modeling. After extraction of the spectral features a Universal Background Model (UBM), i.e., a class-independent GMM with 128 Gaussians, is trained on the whole data set using the Expectation-Maximization (EM) algorithm. The means of the UBM are adapted by relevance *Maximum A Posteriori* (MAP) adaptation in order to get speaker and speech task specific GMMs. The means are then used as speaker- and task-specific features, which form 4992-dimensional (128×39) feature vectors.

3.2. Prosodic Modeling

The prosodic system is not based on any speech recognition output or forced time alignments. Thus, the prosodic features are calculated whenever a voiced speech segment is found. The voiced-unvoiced (VUV) decision is based on the zero crossing rate, the normalized energy of the signal and the maximum energy.

Prosodic base features are calculated on the whole utterance. These are fundamental frequency (F_0), energy, VUV segments, and pitch periods. The structured prosodic features are calculated on the voiced segments. Adjacent segments are merged, when they are separated by less than 50 ms; the corresponding F_0 contour is interpolated to make the segmentation more robust. Context segments, that merge two adjacent segments together, are used additionally. All in all 73 features are calculated for each segment. They model F_0, energy, duration, pauses, jitter, and shimmer. Note that the F_0 features are normalized w.r.t. the mean F_0 and transformed to semitones in order to be comparable across gender. A detailed description of the whole feature set is given in [8]. Finally, we compute mean, minimum, maximum, and standard deviation of these 73 segment features. This forms our 292-dimensional prosodic feature vector.

3.3. Glottal Excitation

The approach estimates the parameters of a physical glottis model. The goal is to find pathology-related changes in the model parameters that reflect voice-related parameters in order to detect speakers suffering from PD. Therefore, the used glottis model should ideally have physically meaningful parameters, in contrast to just describing the shape of the excitation signal. The model should be flexible enough to adequately represent pathology-related changes of the voice.

Considering these requirements we employed the two-mass vocal fold model introduced by Ishizaka and Flanagan [9] and described in Stevens [10] and illustrated in Fig. 3. The model consists of two pairs of masses, larger ones (M_1) representing the inferior part of the vocal folds, and smaller ones (M_2) representing the superior part of the vocal folds. The model is symmetric, i.e., there is no differentiation between the masses of the left and right side. The mechanism depends on the fact that the inferior and superior part of the vocal folds do not move together as a rigid body. There is a certain degree of freedom to move relatively to each other [11]. This freedom is modeled with a coupling compliance by springs (C_c, C_1, C_2). In order to allow a distinct analysis of the excitation signal, the influence of the vocal tract has to be omitted from the speech signal. As an approximation of the excitation signal, the residue of the Linear Predictive Coding (LPC), an inverse filtering of the speech signal with the LPC filter, is calculated in a data-driven optimization procedure. The model parameters are now optimized to match the synthetic excitation signal as close as possible to the LPC residue. Finally, we compute mean, minimum, maximum, and standard deviation of the model parameters and use these values as glottal feature vector.

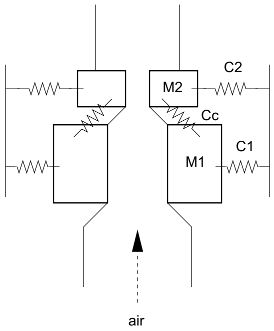

Figure 3: Two-mass vocal fold model by Stevens [10].

3.4. openSMILE

We use the openSMILE toolkit as all-purpose toolkit [12] for feature extraction. We focus on easy to extract acoustic features and use the 1582 acoustic features of the INTERSPEECH 2010 Paralinguistic Challenge baseline [13]. This feature set is obtained by applying a brute-force approach, in which first of all 38 low-level descriptors and their first derivative are computed on the frame level. In a second step, 21 functionals are applied in order to obtain a feature vector of constant length for the whole utterance. Table 1 gives an overview of the low-level descriptors and associated functionals. 16 zero-information features (e. g. the minimum of the fundamental frequency is always zero) are removed from the set of 1596 possible features, and two additional features (*F0 number of onsets* and *turn duration*) are added, resulting in a set of 1582 features.

4. Classification and Experimenal Results

For each of the four systems classification was performed by Support Vector Machines (SVMs) with linear kernel and Sequential Minimal Optimization for learning as implemented in the WEKA toolkit [14]. In preliminary experiments we did not observe any significant differences for various parameter settings, so we decided to use a complexity parameter of 1.0. Evaluation was done in a leave-one-speaker-out (LOO) manner separately for each speech task.

Descriptors	Functionals
PCM loudness	position max./min.
MFCC [0-14]	arithm. mean, std. deviation
log Mel freq. band [0-7]	skewness, kurtosis
LSP frequency [0-7]	lin. regression coeff. 1/2
F0 by sub-harmonic sum.	lin. regression error Q/A
F0 envelope	quartile 1/2/3
voicing probability	quartile range 2−1/3−1/3−2
jitter local	percentile 1/99
jitter DDP	percentile range 99−1
shimmer local	up-level time 75/90

Table 1: Description of the acoustic features based on 38 low-level descriptors and their first derivative and 21 functionals.

Two different sets of experiments were performed with different goals. Detection of PD speakers (Sec. 4.1) is formulated as a two-class classification task between PD and control speakers. Classification of UPDRS (Sec. 4.2) scores among the 88 PD speakers is formulated as a three-class classification problem. For each set of experiments we evaluated the speech-task-dependent recognition results of the stand-alone systems (see Table 2 and 4). Additionally, all nine speech-tasks of one speaker were treated as one audio set (see Table 3 and 5). For each of the four experiments, a system fusion is performed by an unweighted score-level fusion. Probability estimates for each system were obtained by fitting regression models to the output of the SVM [15]. The unweighted mean of these estimates is then used as final class esimate. We intentionally did not use a more sophisticated fusion approach, e.g., fusion by logistic regression, like we did in [16]. For a fair evaluation, this would require a segmentation of the data in order to train the logistic regression component. For each experiment the percentage of correctly classified speakers per single class, and the proportion of **u**nweighted **a**verage recall (UA) is given.

4.1. Detection of Parkinson Speakers – PD vs. Control Speakers

The results of the 2-class problem of detecting whether a speaker is suffering from PD or not are summarized and discussed in this section. Table 2 shows the task-dependent recognition results. The acoustic system achieved the best result (UA of 80.7 %) with task T04, which is a reading task of sentences. However, there was no significant differences between the tasks T01, T02, T04, T05 and T88 for the acoustic system. The prosodic system achieved the best result (73.8 % UA) on the task of repetition of sequences of alternating syllables (T77). This task is motivated by the fact that PD speakers have difficulties in repeating different syllables. The energy/loudness is said to diminish over time and PD speakers tend to speak repeating syllables with variable speech rate [17]. The glottal system achieved its best UA (68.8 %) on task T44 (repetition of syllable /pa/). Different acoustic studies commonly revealed a higher breathiness and harshness for PD speakers [18]. The openSMILE system achieves a UA of 73.9 % on task T88. T88 focuses on a more complex repetition of syllable sequences. A fusion of the different systems did not lead to an improved UA result.

Table 3 shows the results achieved when all recordings tasks are used in combination. This leads to a higher number of available speech data per speaker and results in higher UA results for each stand-alone system. Again, a system fusion did not lead to

system	task	% PD	% CONTROL	% UA
ACOUSTIC	T04	**87.5**	**73.9**	**80.7**
PROS	T77	76.1	71.6	73.8
GLOTTAL	T44	63.6	**73.9**	68.8
OS	T88	77.3	70.5	73.9
FUSION		90.9	65.9	80.0

Table 2: Task-dependent recognition results of the different systems (ACOUSTIC,PROSODIC,GLOTTAL EXCITATION, openSMILE) on the 2-class detection task (PD vs. Control). The last column shows the result after system fusion.

system	% PD	% CONTROL	% UA
ACOUSTIC	**86.4**	**77.3**	**81.9**
PROS	77.3	70.5	73.9
GLOTTAL	72.7	69.3	71.0
OS	78.4	72.7	75.6
FUSION	94.3	63.6	79.0

Table 3: Recognition results of the different systems (ACOUSTIC, PROSODIC, GLOTTAL EXCITATION, openSMILE) on the 2-class detection task (PD vs. Control). All speech tasks have been used in combination. The last column shows the result after system fusion.

improved recognition results.

4.2. Severity of PD: Classification of UPDRS score

The classification of UPDRS scores is addressed as a 3-class task. Table 4 shows the task dependent recognition results of the stand-alone systems and a fusion of theses systems. Again, the best UA result (53.4 %) was achieved by the acoustic system on the text-reading task T02. The prosodic system achieved 39.8 % UA on task T03 (reading question-answer-pairs with stress on certain words). The glottal excitation system achieved 44.4 % UA on the text-reading task. Note that these tasks are the more-complex task, where the focus lies more on speech and articulation rather than on phonation. The openSMILE system achieves an UA of 52.7 % with a more balanced result on the three classes. Fusion of the four systems lead to an improvement (59.1 % UA).

Using all recording tasks in combination did not achieve an improved recognition result (see Table 5). We assume that some of the speaking tasks are suited to discriminate between healthy and PD speakers, but do not help (or even diminish) for a classification of the UPDRS scores.

system	task	% U1	% U2	% U3	% UA
ACOUSTIC	T88	**59.3**	59.4	41.4	53.4
PROS	T03	44.4	40.6	34.5	39.8
GLOTTAL	T02	44.4	40.6	48.3	44.4
OS	T77	51.9	54.5	51.7	52.7
FUSION		44.4	**62.5**	**69.0**	**59.1**

Table 4: Task-dependent recognition results of the different systems (ACOUSTIC, PROSODIC, GLOTTAL EXCITATION, openSMILE) on the 3-class UPDRS task. The last column shows the result after system fusion.

system	% U1	% U2	%U3	% UA
ACOUSTIC	**44.4**	**43.8**	**51.7**	**46.6**
PROS	37.0	43.8	31.0	37.3
GLOTTAL	29.6	37.5	48.3	38.5
OS	44.4	31.3	34.5	36.7
FUSION	29.6	40.6	51.7	40.9

Table 5: Recognition results of the different systems (ACOUSTIC, PROSODIC, GLOTTAL EXCITATION, openSMILE) on the 3-class UPDRS task. All speech tasks have been used in combination. The last column shows the result after system fusion.

5. Concluding Remarks

We did not do any feature selection, due to the fact that we do not have enough data to split into data for feature selection and for classification. For the same reason the fusion of the systems performed by an unweighted score-level fusion that does not require tuning parameters. Thus, our results on system fusion are not conclusive. For the detection task (PD vs. Control) a increase of speech data lead to small improvements for all stand-alone systems. The set-up of the data collection (aquiring only a few minutes of speech) could be used for a cheap screening task. The results, however, are not yet good enough for application in a clinical screening task due to the high number of false alarms.

The 3-class UPDRS-task is a much more difficult problem. The score thresholds were mostly motivated by the fact that we wanted to have a balanced set of classes in terms of number of speakers. Especially the 3rd class spreads across a large range of UPDRS scores. This is worsened by the fact that the medication among the patients was not consistent. For the 3-class UPDRS task an increase of speech data did not lead to recognition improvement. A system fusion of the best stand-alone-systems achieved a significant improvement. This can be explained by the fact that the recognition rates of each feature set across the speech tasks were much less consistent than in the detection task. Thus, one can argue that in the case of categorization of severity the speech task plays an important role. A brute force combination does not help. Still, our results indicate that an appropriate combination of seatures and speech tasks seems a promising approach. Also one has to keep in mind that a surveillance system has several recordings of one speaker and can thus perform a speaker normalization, which should lead to significant improvements.

6. Summary

In this work we focused on the question whether speech carries information about PD being present at a speaker and whether the severity of PD can be classified automatically based on the speech of persons. Four differently motivated system where used for this task. We showed that using all speech tasks in combination lead to an improved recognition result (81.9 % UA) for the detection task (PD present or not). Using the speech data in combination degraded the results on UPDRS classification. But a system fusion on score level achieved an improvement. 59.1 % UA could be reached for this 3-class problem.

7. References

[1] A. E. Lang and A. M. Lozano, "Parkinson's disease," *New England Journal of Medicine*, vol. 339, no. 15, pp. 1044–1053, 1998.

[2] J. A. Logemann, H. B. Fishe, B. Boshes, and E. R. 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, no. 1, pp. 47–57, 1978.

[3] N. Singh, V. Pillay, and Y. E. Choonara, "Advances in the treatment of parkinsons disease," *Progress in Neurobiology*, vol. 81, pp. 29–44, 2007.

[4] A. M. Goberman and C. Coelho, "Acoustic analysis of parkinsonian speech i: Speech characteristics and l-dopa therapy," *Neurorehabilitation*, vol. 17, pp. 237–246, 2002.

[5] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of Persons with Parkinsons Disease by Acoustic, Vocal, and Prosodic Analysis," in *2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, I. S. P. Society, Ed., 2011, pp. 478–483.

[6] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinsons disease," *Journal of the Acoustical Society of America*, vol. 129, pp. 350–367, 2011.

[7] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinsons disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[8] V. Zeissler, "Robuste Erkennung der prosodischen Phänomene und der emotionalen Benutzerzustände in einem multimodalen Dialogsystem," Ph.D. dissertation, University of Erlangen-Nuremberg, Erlangen, Germany, 2012.

[9] K. Ishizaka and J. Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," The Bell System Technical Journal, Tech. Rep., 1972.

[10] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA 02141: The MIT Press, 1998.

[11] G. Fant, *Acoustic Theory of Speech Production*. Netherlands: Mouton, 1960.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[13] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and C. Narayanan, "The Interspeech 2010 Paralinguistic Challenge – Age, Gender, and Affect," in *Proc. Interspeech (2010)*, 2010, pp. 2794–2797.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[15] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., vol. 10. MIT Press, 1998.

[16] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and Gender Recognition Based on Multiple Systems - Early vs. LateFusion," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, ISCA, Ed., 2010, pp. 2830–2833.

[17] A. W. Darkins, V. A. Fromkin, and D. F. Benson, "A Characterization of the Prosodic Loss in Parkinson's Disease," *Brain and Language*, vol. 34, no. 2, pp. 315 – 327, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0093934X88901423

[18] M. K. MacPherson, J. E. Huber, and D. P. Snow, "The Intonation-Syntax Interface in the Speech of Individuals With Parkinson's Disease," *J Speech Lang Hear Res*, vol. 54, no. 1, pp. 19–32, 2011.