I-PROGNOSIS: Verwendung von Sprachmerkmalen als Biomarker zur Detektion der Parkinson-Erkrankung

Hagen Jaeger[1], Michael Stadtschnitzer[1], Alexandra Rizos[2], Fotis Karayiannis[3], George Ntakakis[3] and Leontios Hadjileontiadis[4]

[1] *Fraunhofer IAIS Sankt Augustin, Deutschland*, [2] *King's College London, Hospital Denmark Hill, Vereinigtes Königreich*
[3] *Microsoft Innovation Center Marousi, Griechenland*, [4] *Aristotles-Universität Thessaloniki, Griechenland*

Kurzbeschreibung

Die Parkinson-Krankheit ist eine erbliche Erkrankung des Nervensystems, die sich in motorischen Dysfunktionen durch Neurodegeneration äußert. Durch die im Frühstadium üblicherweise verschleierten Symptome werden Diagnosen häufig erst in einem fortgeschrittenem Krankheitsstadium gestellt. Da die Parkinson-Erkrankung nicht heilbar ist, ist eine Früherkennung wichtig um den Krankheitsverlauf abzumildern und die Lebensqualität des Erkrankten zu verbessern. I-PROGNOSIS ist ein durch das EU-Programm „Horizon 2020" gefördertes Projekt, welches die Früherkennung der Parkinson-Krankheit über anonymisierte Nutzungsaktivitäten von im Alltag verwendeten technischen Geräten erforscht, sowie die Behandlung der Symptome durch geeignete Interventionsmaßnahmen adressiert. Dabei wird auch die Artikulation der Stimme als sensitives Merkmal in die Analyse miteinbezogen.

Einleitung

Eines der ersten frühen Symptome der Parkinson-Erkrankung kann eine abnehmende Leistung der Sprachartikulation sein. Deswegen ist eine Analyse sensitiver Stimm-Merkmale zur frühen Diagnose der Erkrankung Gegenstand aktueller Forschung. Um die Qualität der Aussprache quantifizieren zu können, wird die Extraktion verschiedener Merkmale vorgeschlagen. Einige Ansätze verwenden direkt interpretierbare Dysphonie-Maße wie Jitter, Shimmer und Grundfrequenzstatistiken als sensitive Stimm-Merkmale [1] [2] [3]. Diese Analysen zeigen Erkennungsraten im Bereich von teilweise über 75% bei Verwendung einer Support Vector Machine (SVM) als Klassifikator. Alternativ werden auch nicht direkt interpretierbare Signalstatistiken aus dem Spektrum und Cepstrum beim Übergang von stimmhaften zu stimmlosen lauten (und umgekehrt) zur Analyse gewählt (sog. „Voiced-Unvoiced-Transitions") [4]. Diese Methode konnte auf dem spanischen Testkorpus „PC-GITA" [5] bei Verwendung eines SVM-Klassifikators und Anwendung der Leave-One-Subject-Out (LOSO)-Evaluationsroutine Identifikationsraten von deutlich über 90% zeigen. Spektrale und cepstrale Eigenschaften werden häufig zur Spracherkennung eingesetzt, da man die Vokaltraktfunktion aus ihnen ableiten kann und somit eine hohe Sensitivität für die Charakteristik des stimmerzeugenden Apparates zu erwarten ist [6] [7]. Für die Umsetzung der anonymisierten Merkmalsextraktion innerhalb der i-PROGNOSIS-App wurden die Merkmals-Statistiken der Voiced-Unvoiced-Transitions, sowie der Verlauf der Grundfrequenz ausgewählt, wobei im Folgenden näher auf die Realisierung eingegangen wird.

Umsetzung

Die Implementierung des Algorithmus zur Merkmalsextraktion erfolgte mithilfe des Visual Studio IDE und unter Verwendung von Xamarin in der Programmiersprache C#. Sie wurde für Android-Betriebssysteme implementiert, da dieses eine einfache Umsetzung von Hintergrund-Diensten erlaubt und Benachrichtigungs-Signale bei ein- und ausgehenden Telefonaten anbietet. Abbildung 1 zeigt eine schematische Darstellung der Verarbeitung, die im Folgenden näher erläutert wird.

Sobald ein ankommender oder ausgehender Anruf über das „Off-hook"-Signal (Bildlich: Telefonhörer wird abgehoben) detektiert wird, erfolgt der Start einer Audiosignal-Aufnahme über das geräteinterne Mikrofon. Die Aufnahme endet entweder nach 75 Sekunden, oder durch Auslösen des „On-hook"-Signals (Bildlich: Telefonhörer wird aufgelegt). Das Ergebnis der Aufnahme ist ein unkomprimiertes, diskretes Zeitsignal, welches als Eingangssignal $x_{\mathrm{raw}}(n)$ für die nachfolgende Vorverarbeitung definiert wird. Die Aufnahme erfolgt bei einer Samplingrate von $f_s = 16\,\mathrm{kHz}$, falls diese durch den vom Betriebssystem bereitgestellten Audiorecorder unterstützt wird, alternativ wird mit $f_s = 44,1\,\mathrm{kHz}$ aufgenommen und im Rahmen der Vorverarbeitung ein Resampling nach der Lanczos-Methode mit Blackman-Nuttal-Fensterfunktionskern auf $f_s = 16\,\mathrm{kHz}$ durchgeführt. Des weiteren beinhaltet die Vorverarbeitung eine Entfernung des Gleichstromversatzes durch Anwendung eines IIR-Hochpassfilters 2. Ordnung mit einer Grenzfrequenz von $f_g = 20\,\mathrm{Hz}$ nach Butterworth-Design, woraus sich $x_{\mathrm{HP}}(n)$ als mittelwertbefreites Eingangssignal mit $f_s = 16\,\mathrm{kHz}$ ergibt. Anschließend erfolgt eine Anpassung der Signalaussteuerung, sodass der absolute Maximalwert der Wellenform bei 1.0 liegt. Daraus resultiert

$$x(n) = \frac{x_{\mathrm{HP}}(n)}{\max(|x_{\mathrm{HP}}(n)|)} \tag{1}$$

als vorverarbeitetes Eingangssignal. Das Ziel der nachfolgenden Analyse ist die Erstellung eines abstrakten Merkmalsvektors $\vec{M}_{\mathrm{voice}}$, welcher durch Beobachtung zeitlicher Statistiken der Grundfrequenz des Spektrums, sowie des Cepstrums, zwar sensitive, aber unsensible (oder auch anonymisierte) Informationen über die Stimme enthält.

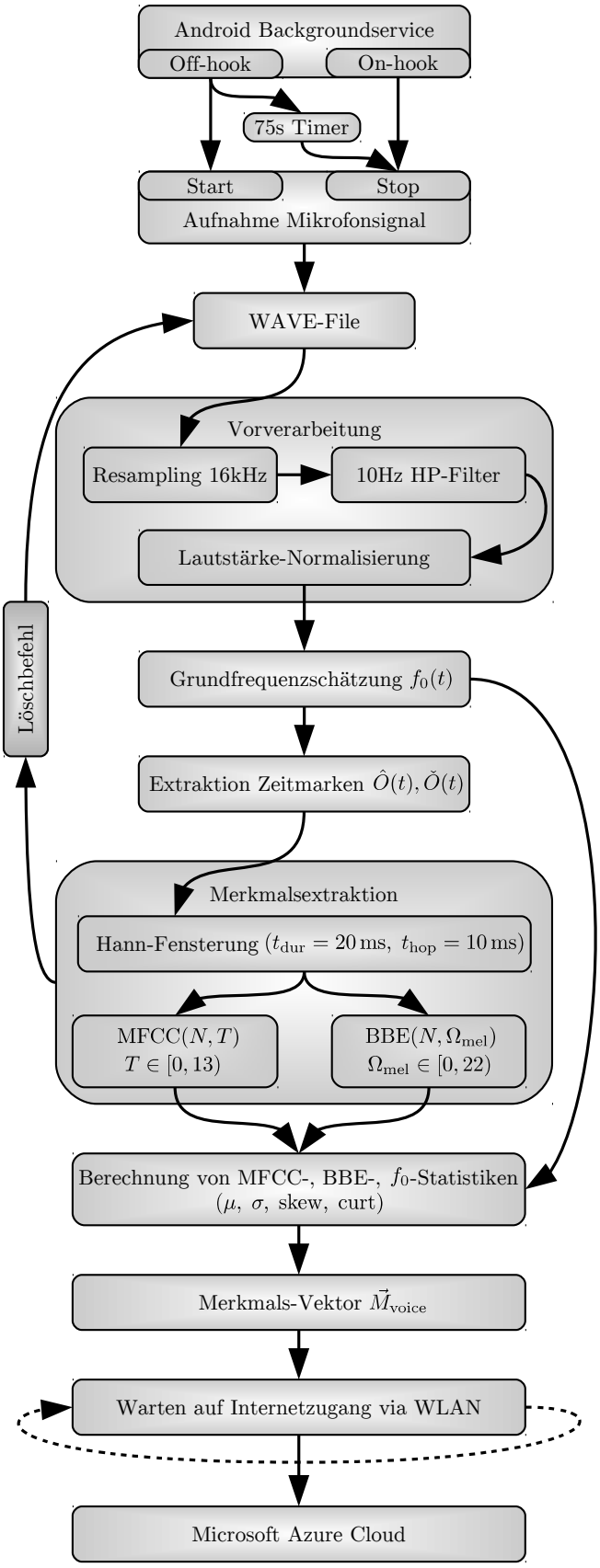


Abbildung 1: Ablaufdiagramm der Extraktion anonymisierter Merkmale für Telefonmitschnitte innerhalb der i-PROGNOSIS-App

Zur Extraktion von Grundfrequenzstatistiken und Identifikation von Voiced-Unvoiced-Transitions, erfolgt eine Schätzung der Grundfrequenz f_0 über die Autokorrelationsmethode innerhalb der Grenzen $f_0^{\min} = 75\,\mathrm{Hz}$ und $f_0^{\max} = 600\,\mathrm{Hz}$. Das Signal wird dazu in Fenster mit Hann-Gewichtung $\mathrm{w}_{\mathrm{hann}}(n)$ der Länge $2L = 640$ Samples ($\widehat{=}\,40\,\mathrm{ms}$) bei Nutzung einer Schrittweite von $\frac{L}{2} = 120$ Samples ($\widehat{=}\,10\,\mathrm{ms}$) aufgeteilt. Die Fenster werden mit ℓ indiziert, sodass

$$x_{f_0}(\underbrace{n - \ell\tfrac{L}{2}}_{n'}, \ell) = x(n) \cdot \mathrm{w}_{\mathrm{hann},2L}(n - \ell\tfrac{L}{2}) \tag{2}$$

als Repräsentation für das Hann-gefensterte Eingangssignal zur Grundfrequenzschätzung geschrieben werden kann. Anschließend wird für jedes Fenster die rechtsseitige Autokorrelationsfunktion mit

$$A_{xx}(m, \ell) = \sum_{m=0}^{L} x_{f_0}(n', \ell) \cdot x_{f_0}(n' - m, \ell) \tag{3}$$

geschätzt. Zur nachfolgenden Grundfrequenzschätzung findet eine Leistungsnormierung statt, sodass

$$||A_{xx}(m, \ell)|| = \frac{A_{xx}(m, \ell)}{A_{xx}(0, \ell)} \tag{4}$$

als normierte Autokorrelationsfunktion geschrieben werden kann. Die Grundfrequenzschätzung wird durch Suche des Maximalwertindex $\hat{m}$ zwischen der unteren und oberen Indexgrenze

$$m_{\mathrm{lo}} = \mathrm{floor}(\frac{f_s}{f_0^{\min}}) \tag{5}$$

$$m_{\mathrm{hi}} = \mathrm{ceil}(\frac{f_s}{f_0^{\max}}) \tag{6}$$

pro Fenster bestimmt, wobei das Maximum eine Mindestamplitude von 0,4 und die Signalblockleistung einen Mindestwert von 0,005 aufweisen muss um als valide Schätzung angenommen zu werden. Somit kann nach Finden des Maximalwertindex mit

$$\hat{m}(\ell) = \mathrm{maxidx}_m\big(\bar{A}_{xx}(m, \ell)\big), \quad \text{mit } m \in [m_{\mathrm{lo}}, m_{\mathrm{hi}}] \tag{7}$$

die Grundfrequenz mittels

$$\begin{aligned} f_0(\ell) &= \frac{f_s}{\hat{m}(\ell)} && \text{wenn} && ||A_{xx}(m_{f_0}(\ell), \ell)|| > 0,4 \\ & && \text{und} && A_{xx}(0, \ell) > 0,005 \\ f_0(\ell) &= 0 && \text{sonst} \end{aligned} \tag{8}$$

geschätzt werden. In Abbildung 2 ist das Ergebnis der Schätzung auf einem zweisekündigen Sprachsegment eines männlichen Sprechers visualisiert, die Grundfrequenzschätzung ist als hervorgehobener weißer Plot eingezeichnet. Basierend auf dieser Schätzung können Zeitmarken für Voiced-Unvoiced-Transitions extrahiert werden, indem jene Stellen, an denen ein Sprung vom oder auf den Wert 0 stattgefunden hat, identifiziert werden.

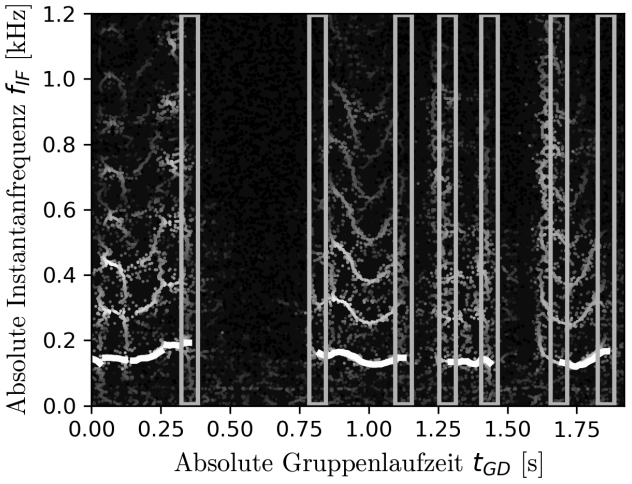


Abbildung 2: Überlagerte Visualisierung einer Grundfrequenz-Schätzung nach der Autokorrelationsmethode (hervorgehobener weißer Plot) und Identifikation von „Voiced-Unvoiced-Transitions“ (in Grau eingezeichnete Kästchen) auf dem „reassigned Spektrogramm“ [8] eines Sprachsignals (männlicher Sprecher).

Die binäre Entscheidung b_{f_0}, die besagt, ob eine Grundfrequenz im betreffenden Signalfenster ℓ detektiert wurde oder nicht, kann mit

$$\begin{aligned} b_{f_0}(\ell) &= 1 \qquad \text{wenn} \qquad f_0(\ell) \neq 0 \\ b_{f_0}(\ell) &= 0 \qquad \text{sonst} \end{aligned} \tag{9}$$

getroffen werden. Die Schätzung kann vereinzelt Fehler, sowie kurze Aussetzer durch plosive Laut-Anteile, oder externe Störungen, enthalten (siehe Abb. 2 bei $t_{\mathrm{GD}} = 0,12\,\mathrm{s}$). Deswegen wird die binäre Entscheidung mit

$$\begin{aligned} \bar{b}_{f_0}(\ell) &= 1 \quad \text{wenn} \quad b_{f_0}(\ell-1) = 1 \quad \text{und } b_{f_0}(\ell) = 0 \\ &\qquad\qquad \text{und } b_{f_0}(\ell+1) = 1 \quad \text{mit } \ell > 0 \\ \bar{b}_{f_0}(\ell) &= 0 \quad \text{wenn} \quad b_{f_0}(\ell-1) = 0 \quad \text{und } b_{f_0}(\ell) = 1 \\ &\qquad\qquad \text{und } b_{f_0}(\ell+1) = 0 \quad \text{mit } \ell > 0 \\ \bar{b}_{f_0}(\ell) &= b_{f_0}(\ell) \quad \text{sonst} \end{aligned} \tag{10}$$

vorab geglättet. Die Zeitmarken T lassen sich anschließend über Änderungen in der binären Entscheidung finden und als

$$T(\ell') = \ell \cdot \tfrac{L}{2} + L \quad \text{wenn} \quad \underbrace{\bar{b}_{f_0}(\ell+1) - \bar{b}_{f_0}(\ell) \neq 0}_{\text{Indexierung durch } \ell'} \tag{11}$$

berechnen. Für die statistische Merkmalsextraktion werden um jede Zeitmarke $T(\ell')$ zentriert sieben Fenster mit Hann-Gewichtung $\mathrm{w}_{\mathrm{hann}}(n)$ der Länge $L = 240$ Samples ($\widehat{=}\,20\,\mathrm{ms}$) bei Nutzung einer Schrittweite von $\frac{L}{2} = 120$ Samples ($\widehat{=}\,10\,\mathrm{ms}$) extrahiert. Das heißt, dass Segmente mit eine Gesamtdauer von jeweils 80ms, zentriert um detektierte Voiced-Unvoiced-Transitions, mit 50% überlappenden Hann-Fenstern analysiert werden. Es wird

$$x(\underbrace{n - \ell\tfrac{L}{2}}_{n'}, \ell') = x(n) \cdot \mathrm{w}_{\mathrm{hann},L}(n - \ell\tfrac{L}{2}), \qquad \underbrace{\ell \in \frac{T}{L} \pm 3}_{\text{Indexierung durch } \ell'} \tag{12}$$

als Eingangssignal für die nachfolgende Merkmalsextraktion definiert. Die spektralen und cepstralen Merkmale werden als Energien in den ersten 22 Bark-Bändern, sowie als die ersten 13 cepstralen Koeffizienten der Mel-Frequenzanalyse berechnet. Um den typischen Leistungsabfall von Sprache zu hohen Frequenzen hin für die Cepstralanalyse zu kompensieren, wird vor der Berechnung ein Filter zur Entfernung der Präemphase angewendet, woraus sich

$$x_e(n', \ell') = \alpha \cdot x(n', \ell') - (1-\alpha) \cdot x_e(n'-1, \ell'), \quad \alpha = 0,03 \tag{13}$$

als Eingangssignal für die Cepstralanalyse ergibt. Dieses wird durch Einführung des diskreten Frequenzindex k und Anwendung von

$$X_e(k, \ell') = \sum_{n'=0}^{L} x_e(n', \ell') \cdot e^{-j2\pi n' \frac{k}{L}} \tag{14}$$

nach Fourier transformiert. Anschließend wird das Leistungs-Spektrum über

$$\tilde{X}_e(k, \ell') = \sqrt{\Re\{X_e(k, \ell')\}^2 + \Im\{X_e(k, \ell')\}^2} \tag{15}$$

bestimmt und mit einer Dreiecks-Filterbank $F_{\mathrm{Mel}}(k, b)$ nach Zwicker [9] in $B_{\mathrm{Mel}} = 26$ Mel-Frequenzbänder mit logarithmischer Amplitude als

$$\tilde{X}^{\log}_{\mathrm{Mel}}(b, \ell') = \log\left(\sum_{k=0}^{L} X_e(k, \ell') \cdot F_{\mathrm{Mel}}(k, b)\right) \tag{16}$$

gruppiert, wobei b den Index des aktuellen Filters definiert. Die cepstralen Koeffizienten der Mel-Frequenz (MFCC), indexiert durch c, ergeben sich aus der diskreten Kosinus-Transformation II der logarithmischen Frequenzbandleistung aus Gleichung 16. Es kann

$$\mathrm{MFCC}(c, \ell') = \sum_{b=0}^{B_{\mathrm{Mel}}} \tilde{X}^{\log}_{\mathrm{Mel}}(b, \ell') \cdot \cos\left(\frac{\pi}{B_{\mathrm{Mel}}} \cdot \left(b + \frac{1}{2}\right) \cdot c\right) \tag{17}$$

als Repräsentation der MFCC-Berechnung auf dem von der Präemphase befreiten Eingangssignal $x_e(n', \ell')$ für alle Signalblöcke ℓ' auf Voiced-Unvoiced-Transitions geschrieben werden. Zusätzlich fließen auch die spektralen Leistungswerte einer Dreiecks-Filterbank $F_{\mathrm{Bark}}(k, b)$ mit Frequenzeinteilung nach Bark [10] (BBE) mit in den Merkmalsvektor ein. Zur Berechnung wird das nicht von der Präemphase befreite Eingangssignal $x(n', \ell')$ nach Gleichung 14 in den Frequenzbereich transformiert und das Leistungs-Spektrum nach Gleichung 15 berechnet (d.h. $x_e(n', \ell')$ und $X_e(k, \ell')$ werden in den Gleichungen durch $x(n', \ell')$ und $\tilde{X}(k, \ell')$ ersetzt), woraus sich $\tilde{X}(k, \ell')$ als resultierendes Leistungs-Spektrum ergibt. Die Energie in Bark-Frequenzbändern kann anschließend mit

$$\tilde{X}_{\mathrm{Bark}}(b, \ell') = \sum_{k=0}^{L} X_e(k, \ell') \cdot F_{\mathrm{Bark}}(k, b) \tag{18}$$

für alle Bänder berechnet werden. Der letzte Schritt der Analyse ist die Berechnung zeitlicher Statistiken auf den

errechneten Merkmalen. Es werden der Mittelwert μ, die Standardabweichung σ, die Schiefe v, sowie die Wölbung ω über alle Blockindizes ℓ (für die Grundfrequenz), bzw. ℓ' (für die MFCC und BBE) berechnet. Die statistischen Berechnungen werden als

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x_i \tag{19}$$

$$\sigma = \frac{1}{N} \sum_{i=0}^{N-1} (x_i - \mu)^2 \tag{20}$$

$$\mathrm{v} = \frac{1}{N} \sum_{i=0}^{N-1} (\frac{x_i - \mu}{\sigma})^3 \tag{21}$$

$$\omega = \frac{1}{N} \sum_{i=0}^{N-1} (\frac{x_i - \mu}{\sigma})^4 \tag{22}$$

definiert, womit der Merkmalsvektor $\vec{M}_{\mathrm{voice}}$ in Vektor-Notation [...] als

$$\begin{aligned}\vec{M}_{\mathrm{voice}} = \Big[&\mu_\ell(f_0), \quad \sigma_\ell(f_0), \quad \mathrm{v}_\ell(f_0), \quad \omega_\ell(f_0), \\ &\mu_{\ell'}(\mathrm{MFCC}),\ \sigma_{\ell'}(\mathrm{MFCC}),\ \mathrm{v}_{\ell'}(\mathrm{MFCC}),\ \omega_{\ell'}(\mathrm{MFCC}) \\ &\mu_{\ell'}(\tilde{X}_{\mathrm{Bark}}), \quad \sigma_{\ell'}(\tilde{X}_{\mathrm{Bark}}), \quad \mathrm{v}_{\ell'}(\tilde{X}_{\mathrm{Bark}}), \quad \omega_{\ell'}(\tilde{X}_{\mathrm{Bark}})\Big]\end{aligned}$$

geschrieben werden kann. Sobald sich das Mobile Endgerät des Benutzers in einem WLAN mit Internetzugang befindet und der Akkuladezustand 20 % nicht unterschreitet, wird der Merkmalsvektor in eine Microsoft Azure Cloud zur Anreicherung hochgeladen und alle temporär auf dem Smartphone gesammelten Daten (Aufnahme und Merkmals-Statistiken) gelöscht.

Zusammenfassung und Ausblick

Die Früherkennung von neurologischen Erkrankungen mithilfe von mobilen Endgeräten kann bei der Frühdiagnose helfen und durch Ergreifung geeigneter Maßnahmen die Lebensqualität der Betroffenen steigern. Das hier vorgestellte Verfahren ist in der Lage anonymisierte Stimm-Merkmale für Übergänge von stimmhaften zu stimmlosen Lauten (und umgekehrt), also zu Zeiten hoher Vokaltrakt- und Glottisvariabilität, zu extrahieren. Bei der Umsetzung wurde sich neben der statistischen Grundfrequenzanalyse an einer modernen Methode [4] orientiert, deren Verwendung statistischer Eigenschaften spektraler und cepstraler Merkmale eine gute Sensitivität für das Erkennen von Änderungen in der Aussprachequalität verspricht. Im weiteren Verlauf des Projektes werden Testdatensätze für unterschiedliche Sprachen aufgenommen, sowie Nutzungsdaten aus dem Alltag der App-Benutzer angereichert, um den Algorithmus anzulernen und seine Leistungsfähigkeit für die Frühdiagnose zu evaluieren. Des weiteren sollen spielerische Interventionsmaßnahmen auf mobilen Endgeräten entwickelt werden, die es erkrankten Personen ermöglichen sollen die einhergehenden Symptome durch aktive Mitarbeit abzumildern (z.B. Training der Verständlichkeit und Stimmartikulation über verschiedene spielerische Anwendungen innerhalb der i-PROGNOSIS-App).

Literatur

[1] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly and P. J. Snyder (**2004**). „Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment“. *Journal of Neurolinguistics*, vol. 17, pp. 439–453.

[2] J. Rusz, R. Cmejla, H. Ruzickova and E. Ruzicka (**2011**). „Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease“. *Journal of the Acoustical Society of America (JASA)*, vol. 129, no. 1, pp. 350–367.

[3] A. Tsanas, M. Little, P. Mcsharry, J. Spielman and L. Ramig (**2012**). „Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease“. *IEEE Transactions on Bio-Medical Engineering*, vol. 59, no. 5, pp. 1264–1271.

[4] J.R Orozco-Arroyave, F. Hönig, J. Arias-Londono, J. Vargas-Bonilla, S. Skodda, J. Rusz and E. Nöth, (**2015**). „Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's Disease“. *Proceedings of 16th Interspeech (INTERSPEECH)*, September 6-10, pp. 95-99, Dresden, Germany.

[5] J.R. Orozco-Arroyave, J.D. Arias-Londoño, J.F. Vargas-Bonilla, M.C. González-Rátiva and E. Nöth (**2014**). „New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease“. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pp. 342-347

[6] R. Vergin, D. O'Shaughnessy and A. Farhat (**1999**). „Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition“. *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525-532

[7] Fawaz S. Al-Anzi and Dia AbuZeina (**2017**). „The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition“. *Kuwait University, Research Project Number EO06/12*.

[8] S. Fulop and K. Fitz (**2006**). „Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications“. *The Journal of the Acoustical Society of America*, vol. 119, pp. 360-371.

[9] E. Zwicker, and E. Terhardt (**1980**). „Analytical expressions for critical-band rate and critical bandwidth as a function of frequency“, *Journal of Acoustical Society of America*, vol. 68, no. 5, pp. 1523-1525.

[10] H. Traunmüller (**1990**). „Analytical Expressions for the Tonotopic Sensory Scale“, *Journal of Acoustical Society of America*, vol. 88, no. 1, pp. 97-100.