University of Tripoli · Faculty of Engineering
Department of Biomedical Engineering · Control and Instrumentation Division

# Voice Signal Analysis Using Machine Learning for Early Detection of Parkinson's Disease

Suhail Mohamed Alhammali (ID 2200208522) · Supervisor: Dr. Adel Adyaf · Spring 2026

| 37 | 74 | 0.822 | 0.701 |
|---|---|---|---|
| subjects studied | acoustic measurements | balanced accuracy | external-corpus AUC |

## 1. The problem

Parkinson's disease damages the part of the brain that controls movement. Its familiar signs are tremor, stiffness, and slowness. What is less widely known is that the disease weakens control of *every* muscle, including the very small muscles used for breathing, for the vocal folds in the throat, and for the tongue and lips.

Speech therefore changes. The voice becomes quieter and flatter in pitch, sometimes breathy or hoarse; consonants become less precise; pauses become more frequent. About nine out of ten patients show some form of this speech disorder, and these changes can begin early — in some documented cases, years before the movement symptoms are clear enough to send a person to a neurologist.

There is no simple laboratory test for Parkinson's disease. Diagnosis depends on a specialist examining movement, and by then much of the damage has already occurred.

## 2. The idea

If the disease changes the voice, and if those changes can be measured, then a voice recording might act as an early warning sign — not a diagnosis, but a reason to see a doctor.

This is attractive because a microphone is cheap, requires no needles or scans, causes no discomfort, and can be used anywhere and repeatedly. The question this project asks is simple: **can a computer measure a voice recording and tell whether its pattern resembles the voices of people with Parkinson's disease or the voices of healthy people?**

## 3. How the system works

The system takes an ordinary audio file and passes it through four stages.

**First, it cleans the recording** in a standard way: one audio channel, one fixed quality setting, silence removed from the beginning and end, and the volume levelled so that a loud recording and a quiet one can be compared fairly. One thing is deliberately *not* done: noise removal. Noise-removal software works by smoothing away irregularities in a signal — and small irregularities in the voice are exactly the medical evidence being looked for. Cleaning the sound too aggressively would erase the very thing being measured.

**Second, it cuts the recording into ten-second pieces.** Measuring many short pieces and averaging them is more reliable than measuring a whole recording once, in the same way that taking a person's blood pressure several times gives a truer picture than a single reading.

**Third, it measures the voice.** From each piece the system computes 74 numbers. These are not arbitrary: each one is chosen because it corresponds to something physical in the throat and mouth. Some describe how high or low the voice is and how much it varies — flat, monotone speech is a classic sign of the disease. Some describe how regularly the vocal folds vibrate, since unstable muscle control produces uneven vibration. Some describe how much of the voice is clear tone and how much is breathy air, since weakened muscles do not close the vocal folds completely. Others describe the rhythm of pauses, and the shape and speed of movement of the mouth and tongue.

Building this measurement stage was the main engineering work of the project. No ready-made set of measurements was downloaded; all of it was written and tested as part of this work.

**Fourth, it learns and judges.** The 74 numbers from many people, with their known group, are given to a machine learning model, which finds the patterns that separate the two groups. The model used is a deliberately simple and transparent one, so that it can be asked afterwards *which measurements it relied on* — an answer that a modern "deep learning" system could not give.

## 4. The central question: how do you test it honestly?

This is where most of the project's effort went, and it is its main scientific contribution.

A model must be tested on people it has never seen. If it is tested on people it was trained on, a high score proves nothing — like a student who is given the exam questions in advance.

The dataset used here contains two recordings from each person. The danger is subtle: if one recording of a person is used for training and the other for testing, the model can succeed simply by *recognising that person's voice*, rather than by detecting the disease. It would then look excellent on this data and fail completely on anyone new. This is called data leakage.

The project prevents it strictly: every recording of a person always stays on the same side of the divide, and the program contains a check that stops it immediately if this rule is ever broken.

The size of the problem was then *measured*, not merely asserted. The same system, the same model, the same data were run twice, changing only how the data was divided. Divided correctly, the result was 0.775. Divided incorrectly, it was 0.909. That difference is not skill; it is memorisation. This single comparison explains why many published studies in this field report accuracies of 95% or more, and why this project's lower number is the more trustworthy one.

## 5. What was found

Working with recordings from 37 people, the final system judges a person correctly about 82% of the time — measured only on people it had never encountered. When asked which measurement mattered most, it answered: the range of pitch variation. That is precisely the flat, monotone speech that doctors have described as a hallmark of the disease for decades. A system built without any medical knowledge rediscovered a known clinical sign.

The system was then given a much harder test. It was frozen — no retraining, no adjustment — and applied to a completely different collection of recordings from Italy: different language, different microphones, different rooms. Performance dropped, as expected, but did not collapse: given one patient and one healthy person at random, the system still ranked the patient as more likely about 70% of the time. Some genuine, transferable information about the voice is therefore being captured, and the cost of changing conditions was measured rather than guessed.

Two decisions in the project are worth noting because they went against the temptation to report a bigger number. First, a more complicated combination of models scored slightly higher, and was rejected: the rules for accepting an improvement had been fixed in advance, and the difference was too small to be distinguished from chance. Second, the programme reports "inconclusive" instead of an answer whenever the evidence is weak. In the Italian test, this meant that *every single* healthy elderly speaker received "I cannot tell" rather than a confident wrong label.

## 6. What this system is not

It does not diagnose Parkinson's disease, and it is not designed to. Many other things change a voice — a cold, a sore throat, ageing, smoking, tiredness. The system reports only that a recording's acoustic pattern resembles one group in a research dataset more than the other.

Its limits are stated openly: 37 people is a small number, all recorded with one device in one language; the system has never been tested on people in the earliest stage of the disease specifically; and its score is not a probability of illness.

The honest summary is this. Measurable properties of speech do differ between people with Parkinson's disease and healthy people. A simple, explainable model can use those differences with roughly 82% accuracy on people it has never heard, and roughly 70% of that ability survives a move to another country and language. And a responsible system, faced with a voice unlike anything it was trained on, should say so rather than guess.

# تحليل إشارة الصوت باستخدام تعلّم الآلة للكشف المبكر عن مرض باركنسون

سهيل محمد الحمالي، رقم القيد 2200208522، بإشراف الدكتور عادل أضياف

شعبة التحكم والقياس، قسم الهندسة الطبية الحيوية، كلية الهندسة، جامعة طرابلس، فصل ربيع 2026

## 1. المشكلة

يُتلف مرض باركنسون الجزء من الدماغ المسؤول عن التحكم بالحركة. وأعراضه المعروفة هي الرعاش والتيبّس وبطء الحركة. لكن الأقل شهرة أن المرض يُضعف التحكم في **كل** عضلة في الجسم، بما فيها العضلات الصغيرة جداً المستخدمة في التنفّس، وفي تحريك الحبال الصوتية داخل الحنجرة، وفي تحريك اللسان والشفتين.

ولذلك يتغيّر الكلام. فيصبح الصوت أخفض وأكثر رتابة في النبرة، وقد يصير مهموساً أو مبحوحاً، وتصير الحروف الساكنة أقل دقة، وتكثر التوقّفات. ويظهر شكل من أشكال هذا الاضطراب لدى نحو تسعة من كل عشرة مرضى. وقد تبدأ هذه التغيّرات مبكراً، بل رُصدت في حالات موثّقة قبل سنوات من وضوح الأعراض الحركية بما يكفي لدفع المريض إلى زيارة طبيب أعصاب.

ولا يوجد فحص مخبري بسيط لمرض باركنسون. فالتشخيص يعتمد على فحص مختصّ للحركة، وعند ذلك يكون جزء كبير من الضرر قد وقع بالفعل.

## 2. الفكرة

إذا كان المرض يغيّر الصوت، وإذا أمكن قياس هذه التغيّرات، فقد يصلح تسجيل صوتي بوصفه علامة إنذار مبكّر. لا تشخيصاً، بل سبباً لزيارة الطبيب.

وهذه الفكرة جذابة لأن الميكروفون رخيص، ولا يحتاج إبراً ولا أشعة، ولا يسبّب أي ألم، ويمكن استخدامه في أي مكان وتكراره متى شئنا. والسؤال الذي يطرحه هذا المشروع بسيط: **هل يستطيع الحاسوب أن يقيس تسجيلاً صوتياً ويقول إن نمطه يشبه أصوات المصابين بمرض باركنسون أم يشبه أصوات الأشخاص الأصحّاء؟**

## 3. كيف يعمل النظام

يأخذ النظام ملفاً صوتياً عادياً ويمرّره بأربع مراحل.

**أولاً، ينظّف التسجيل** بطريقة موحّدة: قناة صوتية واحدة، وإعداد جودة ثابت، وحذف الصمت من بداية التسجيل ونهايته، وتسوية مستوى الصوت حتى يمكن مقارنة تسجيل مرتفع بتسجيل منخفض مقارنة عادلة. وهناك أمر واحد **لا** يُفعل عمداً: إزالة الضجيج. فبرامج إزالة الضجيج تعمل بتنعيم عدم الانتظام في الإشارة، وعدم الانتظام الصغير في الصوت هو بالضبط الدليل الطبي الذي نبحث عنه. فالتنظيف المبالغ فيه يمحو الشيء الذي نريد قياسه.

**ثانياً، يقطّع التسجيل إلى قطع مدّتها عشر ثوانٍ.** فقياس عدة قطع قصيرة ثم أخذ متوسّطها أوثق من قياس التسجيل كاملاً مرة واحدة، تماماً كما أن قياس ضغط دم شخص عدة مرات يعطي صورة أصدق من قياس واحد.

**ثالثاً، يقيس الصوت.** يحسب النظام من كل قطعة أربعة وسبعين رقماً. وهذه الأرقام ليست اعتباطية، بل اختير كل واحد منها لأنه يقابل شيئاً مادياً يحدث في الحنجرة والفم. فبعضها يصف مدى ارتفاع الصوت أو انخفاضه ومقدار تغيّره، لأن الكلام المسطّح الرتيب علامة كلاسيكية للمرض. وبعضها يصف مدى انتظام اهتزاز الحبال الصوتية، لأن التحكم العضلي غير المستقر ينتج اهتزازاً غير منتظم. وبعضها يصف كم من الصوت نغمة صافية وكم منه هواء مهموس، لأن العضلات الضعيفة لا تُغلق الحبال الصوتية إغلاقاً كاملاً. وبعضها الآخر يصف إيقاع التوقّفات، وشكل حركة الفم واللسان وسرعتها.

وبناء مرحلة القياس هذه هو العمل الهندسي الأساسي في المشروع. فلم تُحمَّل أي مجموعة قياسات جاهزة، بل كُتبت كلها واختُبرت ضمن هذا العمل.

**رابعاً، يتعلّم ويحكم.** تُعطى الأرقام الأربعة والسبعون الخاصة بأشخاص كثيرين، مع معرفة المجموعة التي ينتمي إليها كل منهم، إلى نموذج تعلّم آلة، فيكتشف الأنماط التي تفرّق بين المجموعتين. والنموذج المستخدم بسيط وشفاف عن قصد، حتى يمكن أن نسأله بعد ذلك: **على أي قياسات اعتمدت؟** وهو سؤال لا يستطيع نظام «التعلّم العميق» الحديث أن يجيب عنه.

## 4. السؤال المركزي: كيف نختبره بأمانة؟

هنا ذهب معظم جهد المشروع، وهذه هي مساهمته العلمية الأساسية.

يجب أن يُختبر أي نموذج على أشخاص لم يرهم من قبل. فإن اختُبر على الأشخاص أنفسهم الذين تدرّب عليهم، لم تُثبت الدرجة العالية شيئاً، تماماً كطالب أُعطي أسئلة الامتحان قبل الامتحان.

وتحتوي قاعدة البيانات المستخدمة هنا على تسجيلين لكل شخص. والخطر دقيق: إذا استُخدم أحد تسجيلَي شخص ما في التدريب والآخر في الاختبار، فقد ينجح النموذج لمجرّد أنه **تعرّف على صوت ذلك الشخص**، لا لأنه كشف المرض. وعندها يبدو ممتازاً على هذه البيانات ويفشل تماماً مع أي شخص جديد. ويُسمّى هذا تسريب البيانات.

ويمنعه المشروع منعاً صارماً: فكل تسجيلات الشخص الواحد تبقى دائماً على جانب واحد من التقسيم، ويحتوي البرنامج على فحص يوقفه فوراً إن خُرقت هذه القاعدة.

ثم **قِيس** حجم المشكلة بدل الاكتفاء بادّعائها. فشُغّل النظام نفسه والنموذج نفسه والبيانات نفسها مرتين، ولم يتغيّر سوى طريقة تقسيم البيانات. فكانت النتيجة بالتقسيم الصحيح 0.775، وبالتقسيم الخاطئ 0.909. وهذا الفرق ليس مهارة، بل حفظ. وهذه المقارنة وحدها تفسّر سبب إعلان كثير من

الدراسات المنشورة في هذا المجال دقّة تبلغ خمسة وتسعين بالمئة أو أكثر، وسبب كون الرقم الأقل في هذا المشروع هو الأجدر بالثقة.

## 5. ما الذي وُجد

بالعمل على تسجيلات سبعة وثلاثين شخصاً، يحكم النظام النهائي على الشخص حكماً صحيحاً في نحو اثنين وثمانين بالمئة من الحالات، مقيسة على أشخاص لم يصادفهم قط. وحين سُئل عن أهم قياس اعتمد عليه، كان جوابه: مدى تغيّر النبرة. وهو بالضبط الكلام المسطّح الرتيب الذي يصفه الأطباء علامةً مميّزة للمرض منذ عقود. أي أن نظاماً بُني بلا أي معرفة طبية أعاد اكتشاف علامة سريرية معروفة.

ثم خضع النظام لاختبار أصعب بكثير. فقد جُمّد، بلا إعادة تدريب وبلا أي تعديل، وطُبّق على مجموعة تسجيلات مختلفة تماماً من إيطاليا: لغة مختلفة، وميكروفونات مختلفة، وغرف مختلفة. فانخفض الأداء كما هو متوقّع، لكنه لم ينهَر: فإذا أُخذ مريض وشخص سليم عشوائياً، ظلّ النظام يرتّب المريض بوصفه الأرجح في نحو سبعين بالمئة من الحالات. أي أن هناك معلومة حقيقية عن الصوت قابلة للانتقال، وأن كلفة تغيّر الظروف قِيست بدل أن تُخمَّن.

وهناك قراران في المشروع يستحقّان الذكر لأنهما خالفا إغراء إعلان رقم أكبر. الأول أن تركيبة أعقد من النماذج حقّقت نتيجة أعلى قليلاً، فرُفضت: لأن قواعد قبول أي تحسين كانت قد ثُبِّتت مسبقاً، ولأن الفرق كان أصغر من أن يُميَّز عن المصادفة. والثاني أن البرنامج يعلن «غير حاسم» بدل إعطاء جواب كلما كان الدليل ضعيفاً. وفي الاختبار الإيطالي كان معنى ذلك أن **كل** متحدث سليم كبير في السن، بلا استثناء واحد، حصل على «لا أستطيع الجزم» بدل تصنيف خاطئ واثق.

## 6. ما ليس هذا النظام

إنه لا يشخّص مرض باركنسون، وليس مصمّماً لذلك. فأشياء كثيرة أخرى تغيّر الصوت: زكام، أو التهاب في الحلق، أو تقدّم في العمر، أو تدخين، أو إرهاق. وكل ما يُبلّغ عنه النظام هو أن النمط الصوتي لتسجيل ما يشبه إحدى مجموعتَي قاعدة بيانات بحثية أكثر من الأخرى.

وحدوده معلنة بصراحة: سبعة وثلاثون شخصاً عدد صغير، وكلهم سُجّلوا بجهاز واحد وبلغة واحدة، والنظام لم يُختبر تحديداً على أشخاص في المرحلة الأبكر من المرض، ودرجته ليست احتمالاً للإصابة.

والخلاصة الأمينة هي التالية. توجد فعلاً خصائص قابلة للقياس في الكلام تختلف بين المصابين بمرض باركنسون والأصحّاء. ويستطيع نموذج بسيط وقابل للتفسير أن يستفيد من هذه الفروق بدقة تقارب اثنين وثمانين بالمئة على أشخاص لم يسمعهم من قبل، وأن نحو سبعين بالمئة من هذه القدرة تصمد عند الانتقال إلى بلد ولغة أخريين. وأن النظام المسؤول، حين يواجه صوتاً لا يشبه شيئاً مما تدرّب عليه، ينبغي أن يقول ذلك بدل أن يخمّن.