Uso de espectrogramas y estimación de incertidumbre en la detección de la enfermedad de Parkinson a partir de voz

Wilberth Ferney Córdoba Canchala

1. Contexto de aplicación

La enfermedad de Parkinson (EP) es un trastorno neurodegenerativo progresivo que afecta el control motor y la producción del habla. Uno de los síntomas tempranos es la disartria, que se refleja en cambios en la voz como la reducción de la intensidad, la monotonía y la dificultad en la articulación de fonemas. Aunque existen modelos de machine learning aplicados a voz, a menudo se centran únicamente en la clasificación binaria (Parkinson vs sano) sin considerar la **incertidumbre** de sus predicciones(Kendall & Gal, 2017).

Supongamos que entrenamos una red neuronal convolucional (CNN) para clasificar entre voces sanas y voces con Parkinson. El modelo puede dar un resultado como: "Probabilidad de Parkinson: 95%". A primera vista, esto parece un diagnóstico seguro. Sin embargo, esa cifra no necesariamente significa que el modelo esté confiado de forma realista, sino que corresponde a la salida de una función *softmax*.

Aquí es donde aparecen los **atajos de aprendizaje** o **correlaciones espurias**(Arias-Londoño & Godino-Llorente, 2024). Por ejemplo, si en el dataset de voces con Parkinson la mayoría de grabaciones se hicieron en un consultorio con eco, y las voces sanas en ambientes silenciosos, la red podría estar diferenciando el ruido de fondo y no la patología.

En este sentido, el uso de incertidumbre mediante técnicas como *Monte Carlo Dropout* (Gal & Ghahramani, s. f.) permite detectar cuándo el modelo está inseguro, e incluso visualizar esas zonas con mapas de atención (Grad-CAM)(Arias-Londoño & Godino-Llorente, 2024).

Para llevar estas ideas a un caso práctico, necesitamos una representación adecuada de la voz. Los **espectrogramas** permiten capturar información en el dominio temporal y frecuencial, lo cual es clave para identificar alteraciones sutiles del habla que no se aprecian en el audio (Guerrero-López et al., 2024).

2. Objetivo de Machine Learning

En este proyecto usaremos una **CNN base, de arquitectura simple y pocas capas**, como un prototipo rápido para observar cómo evoluciona la incertidumbre en diferentes fases del entrenamiento. No buscamos aún un modelo clínico definitivo, sino un ejercicio didáctico y acotado, esto se alinea con la recomendación de

proyectos de bajo alcance computacional, sirviendo como base para posteriores ampliaciones con más datos y arquitecturas más complejas a medida que avance mi investigación de doctorado.

1. Entrenar una CNN ligera sobre espectrogramas.
2. Evaluar no solo la precisión del modelo, sino también la distribución de incertidumbre en sus predicciones.
3. Visualizar, mediante **Class Activation Maps (CAM)**, las regiones del espectrograma donde el modelo presenta mayor atención o incertidumbre.
4. Analizar cómo cambia la incertidumbre en diferentes fases del entrenamiento.

3. Dataset

Se utilizará, un subconjunto de la *Saarbrücken Voice Database (SVD)* —carpeta Morbus Parkinson—, centrado en vocales sostenidas. Nos basaremos explícitamente en el pipeline de preprocesamiento descrito por (Ibarra et al., 2023)

Tipo de datos. Grabaciones de voz (.nsp) con vocales sostenidas (/a/, /i/, /u/).

Tamaño real usado. 26 grabaciones únicas: 13 de pacientes con Parkinson (subcarpeta *Morbus Parkinson*) y 13 de sujetos sanos (subcarpeta *Healthy*).

Data Augmentation: Dado que el número de muestras es limitado, se aplicarán técnicas de **aumento de datos** para incrementar la variabilidad y robustez del modelo, siguiendo un enfoque inspirado en (Ibarra et al., 2023), aunque adaptado a vocales sostenidas. Entre las técnicas consideradas están: Pitch shifting, Time stretching, Adición de ruido blanco o ambiental, SpecAugment, enmascaramiento aleatorio de bandas de frecuencia y tiempo en el espectrograma.

Clases. Binaria: **Parkinson vs saludable**.

Tamaño en disco. ~1.6 MB (solo los .nsp).

Siguiendo el pipeline (Ibarra et al., 2023) se aplicará resampleo a 44.1 kHz, Conversión a espectrogramas de Mel: 65 bandas, hop de 10 ms, amplitud en dB, Normalización z-score para estandarizar los espectrogramas (~65×41 píxeles por segmento). Dada la baja cantidad de audios, se aplicará validación cruzada o *hold-out* minimalista para evaluar desempeño.

4. Métricas de desempeño

De acuerdo a (Ibarra et al., 2023), las métricas recomendadas son: Accuracy (precisión de clasificación), F1-score (balance entre sensibilidad y especificidad). De

acuerdo a (Kendall & Gal, 2017), se utilizará la Calibración de incertidumbre (ECE – Expected Calibration Error). En cuanto a las métricas de aplicabilidad, buscamos la capacidad de identificar muestras con alta incertidumbre, lo cual puede indicar que el modelo requiere más datos o que el caso debe revisarse con mayor cuidado y la Visualización de atención mediante CAM, como herramienta de apoyo al especialista.

5. **Referencias y resultados previos**

Arias-Londoño, J. D., & Godino-Llorente, J. I. (2024). Analysis of the Clever Hans effect in COVID-19 detection using Chest X-Ray images and Bayesian Deep Learning. *Biomedical Signal Processing and Control*, *90*, 105831. https://doi.org/10.1016/j.bspc.2023.105831

Gal, Y., & Ghahramani, Z. (s. f.). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*.

Guerrero-López, A., Arias-Londoño, J. D., Shattuck-Hufnagel, S., & Godino-Llorente, J. I. (2024). *MARTA: A model for the automatic phonemic grouping of the parkinsonian speech*. Preprints. https://doi.org/10.36227/techrxiv.171084943.31044695/v1

Ibarra, E. J., Arias-Londoño, J. D., Zañartu, M., & Godino-Llorente, J. I. (2023). Towards a Corpus (and Language)-Independent Screening of Parkinson's Disease from Voice and Speech through Domain Adaptation. *Bioengineering*, *10*(11), 1316. https://doi.org/10.3390/bioengineering10111316

Kendall, A., & Gal, Y. (2017). *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?* (No. arXiv:1703.04977). arXiv. https://doi.org/10.48550/arXiv.1703.04977