

# **ANÁLISIS DE LA INFLUENCIA DE CARACTERÍSTICAS VOCALES EN MODELOS DE DETECCIÓN DE VOCES CLONADAS MEDIANTE IA**

**Fernando Mateos**



# Índice

- |          |  |          |  |          |  |
|----------|--|----------|--|----------|--|
| <b>1</b> | <b>Introducción: Problemas y Objetivos</b> | <b>2</b> | <b>Historia de la clonación de voz</b> | <b>3</b> | <b>Procesamiento de Audio</b>          |
| <b>4</b> | <b>Características Acústicas</b>           | <b>5</b> | <b>Arquitectura de la red</b>          | <b>6</b> | <b>Entrenamiento del modelo</b>        |
| <b>7</b> | <b>Análisis y evaluación</b>               | <b>8</b> | <b>Resultados y conclusión</b>         | <b>9</b> | <b>Mejoras y Posibles Aplicaciones</b> |



# Introducción: Problemas y Objetivos

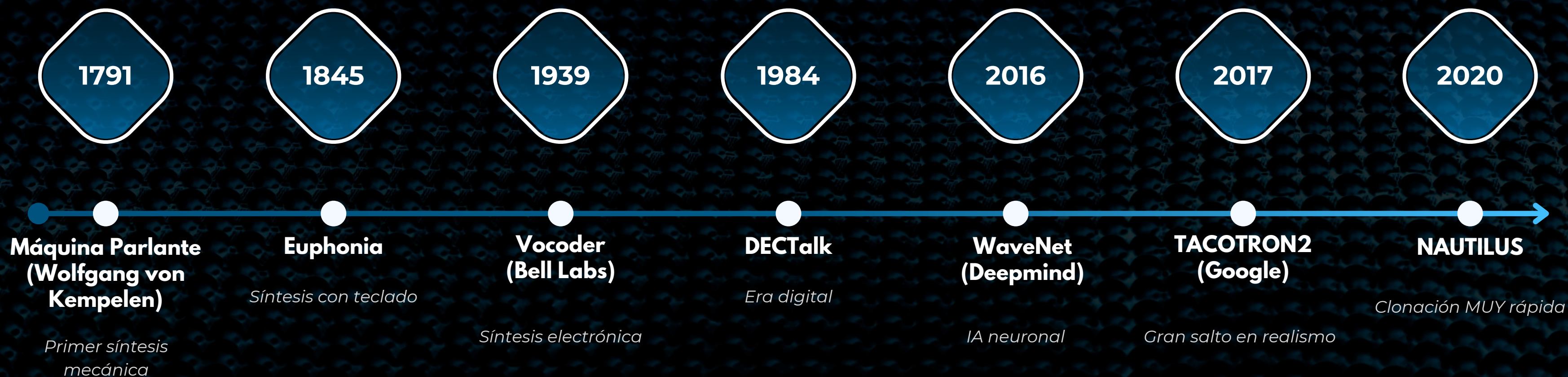
**La clonación de voces empieza a ser una amenaza en auge y existen ya casos muy graves de deepfakes (Hong Kong 24mill.)**

**Las voces clonadas pueden afectar la privacidad y la seguridad. La capacidad de replicar voces con alta fidelidad permite suplantar identidades, cometer fraudes y difundir desinformación**

- **El objetivo principal es determinar si existen o no diferencias estadísticamente significativas en las métricas de evaluación para las diferentes características vocales**
- **Obtener un conjunto de características que nos permitan distinguir las voces de manera eficaz y que marquen la diferencia estadística**



# Historia de la Clonación de voz





# Procesamiento de Audio

## Digitalización:

La señal de voz analógica se convierte en datos digitales mediante muestreo y cuantificación, permitiendo su análisis automático

## Preprocesamiento:

Se normaliza el volumen, se filtra el ruido y se eliminan silencios para mejorar la calidad y homogeneidad de las grabaciones

## Extracción de características:

Se generan espectrogramas y MFCCs, que resumen la información acústica clave para entrenar modelos de detección de voces clonadas



# Características Acústicas Utilizadas

1

**MFCCs:** Coeficientes que representan una señal de audio

2

**RSM:** Intensidad media y energía de la voz

3

**HNR:** Relación entre armónicos y ruido vocal.

4

**Formantes:** Picos de resonancia del tracto vocal

5

**Contraste Espectral:** Diferencia entre picos y valles espectrales

6

**Tasa de cruces por 0:** Frecuencia de cambios de signo en la señal

7

**MFCCs + RMS**

8

**MFCCs + HNR**

9

**BASICO**  
(MFCCs+HNR+RMS)

10

**CALIDAD VOCAL**  
(MFCCs+HNR+FORMANTES)

11

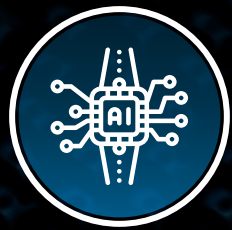
**ESPECTRAL AVANZADO**  
(MFCCs+CONTRASTE ESPECTRAL+ANCHO DE BANDA)

12

**ANÁLISIS COMPLETO**  
(TODOS+ MFCCs delta)

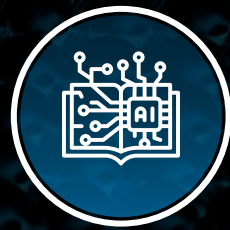


# Arquitectura de la red



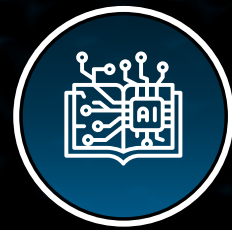
## Arquitectura Simple

Utilizo una arquitectura simple (CNN 1D) para que los resultados de las características no se camuflen en redes muy complejas o especializadas



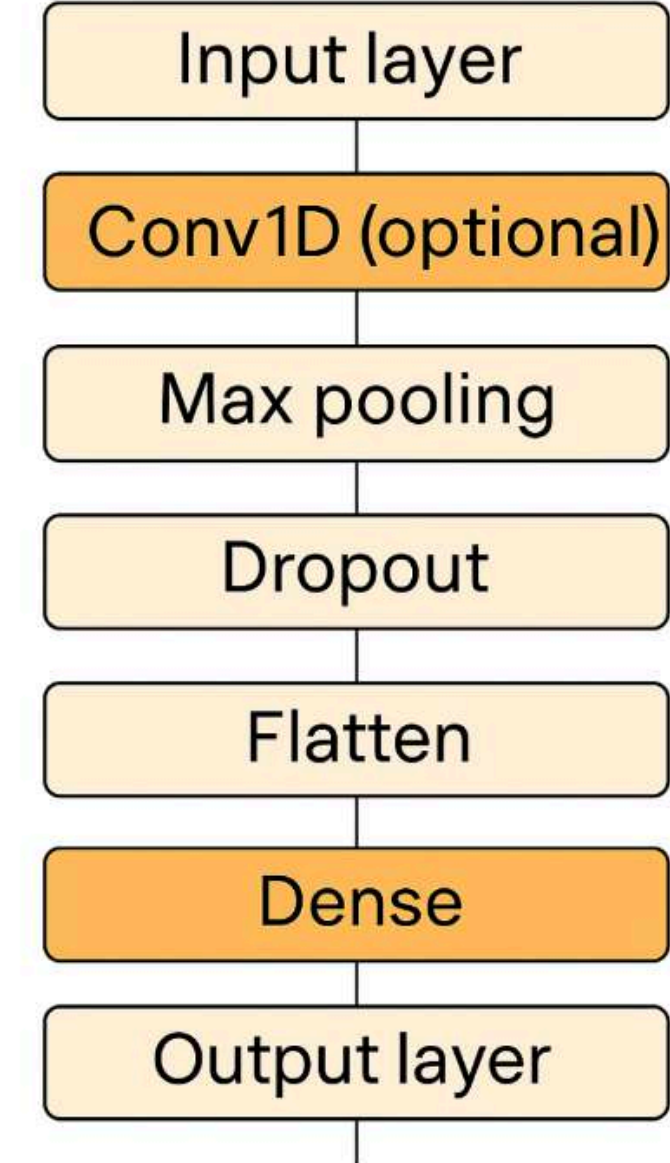
## Hiperparámetros

La arquitectura es flexible, pues se pueden escoger algunos hiperparámetros cómo la tasa de DROPOUT, el n° de filtros en la capa conv...



## Eficiencia

La arquitectura nos permite poder ejecutar los entrenamientos y test en dispositivos con características “normales”



Esquema de la arquitectura de la red





# Entrenamiento del modelo

---

Se cargan los audios reales y clonados, extrayendo las características acústicas elegidas. El dataset está preprocesado y normalizado

---

Los datos se dividen en conjuntos de entrenamiento (80%) y validación (20%), manteniendo el equilibrio entre clases. Tener suficientes datos de entrenamiento y detectar posibles problemas de sobre ajuste

---

Arquitectura adaptativa, añadiendo capas convolucionales solo si el número de características lo requiere. Función de pérdida binaria y el optimizador Adam, monitorizando el rendimiento para aplicar parada temprana y evitar el sobre-entrenamiento

---

Finalmente, el modelo ajusta sus pesos a lo largo de varias épocas, procesando los datos en lotes y almacenando las métricas clave para el análisis

---

Finalmente, el modelo ajusta sus pesos a lo largo de varias épocas, procesando los datos en lotes y almacenando las métricas clave para el análisis



# Análisis y evaluación

Entrenamiento y test  
(12 CARACTERÍSTICAS  
Y 20 ARQUITECTURAS)

Precision  
Sensibilidad (Recall)  
Especificidad (Specificity)  
Exactitud (Accuracy)  
F1-Score  
Área bajo AUC-ROC

Se usa Kruskal-Wallis  
para la comparación y  
se hacen test post-hoc  
Mann-Whitney con  
corrección de  
Bonferroni

Primero se verifica la  
normalidad de los  
datos (Shapiro-Wilk).  
En ningún resultado  
había normalidad

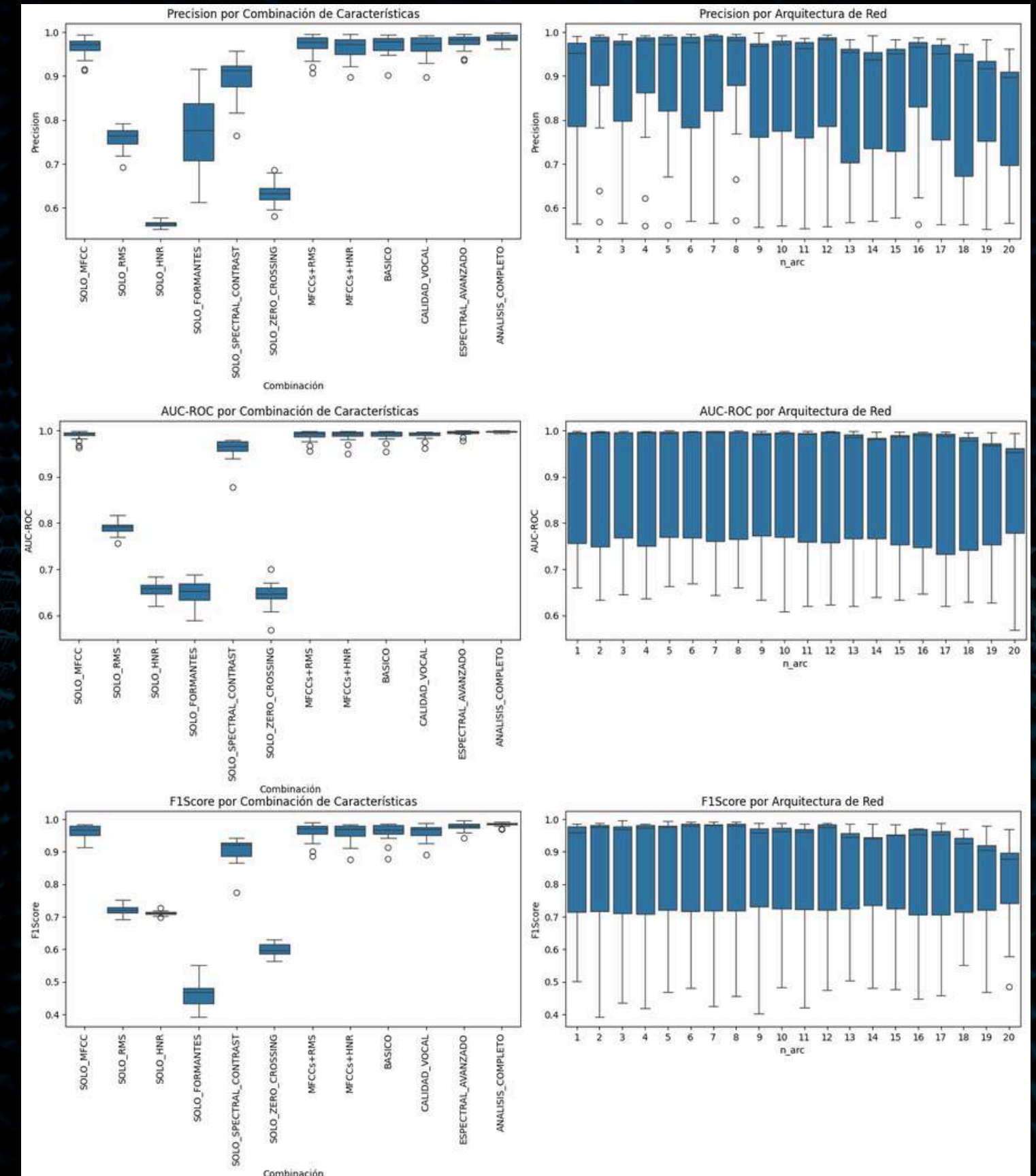
Por último se calcula  
el índice de  
dominancia entre la  
combinación y la  
arquitectura



# Resultados

Para las 6 métricas analizadas se obtienen resultados muy similares

- No existe normalidad en los datos
- No se ven cambios importantes en las diferentes arquitecturas pero sí entre las características
- MFCCs y Contraste Espectral destacan por separado
- Todas las combinaciones que usan los MFccs tienen los mejores resultados





# Resultados

Los test estadísticos son claros y coinciden para TODAS las métricas

- Entre el 53 y el 70 por ciento de las comparaciones post-hoc para TODAS las métricas y con los 3 porcentajes de audios son significativas
- Todos los H-Estadísticos son muy grandes lo que indica que las diferencias entre las medianas de los grupos son significativas
- Los p-value cercanos a 0 (casi iguales a 0) indican que se puede rechazar la hipótesis nula de que no existen diferencias estadísticamente significativas entre los grupos analizados

## 6.3.1. Precision

Comparación entre porcentajes de datos

Cuadro 6.1: Análisis estadístico de Precision por porcentaje de datos

Porcentaje	H-estadístico	$\epsilon_{Comb}^2$	$\epsilon_{Arc}^2$	Índice D	Significancia
1 %	175.98	0.7236	0.0183	39.54	p <0.001
5 %	191.38	0.7911	0.0283	27.99	p <0.001
10 %	204.12	0.8145	0.0245	33.24	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.2: Comparaciones Mann-Whitney significativas para Precision

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	45	68.2 %
5 %	66	42	63.6 %
10 %	66	38	57.6 %

## 6.3.6. AUC-ROC

Comparación entre porcentajes de datos

El AUC-ROC presenta los valores más altos de dominancia, siendo la métrica que mejor evidencia la superioridad de las características acústicas.

Cuadro 6.11: Análisis estadístico de AUC-ROC por porcentaje de datos

Porcentaje	H-estadístico	$\epsilon_{Comb}^2$	$\epsilon_{Arc}^2$	Índice D	Significancia
1 %	178.96	0.7389	0.0145	50.96	p <0.001
5 %	204.75	0.8357	0.0189	44.21	p <0.001
10 %	215.23	0.8634	0.0198	43.61	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.12: Comparaciones Mann-Whitney significativas para AUC-ROC

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	47	71.2 %
5 %	66	45	68.2 %
10 %	66	41	62.1 %

## 6.3.5. F1-Score

Comparación entre porcentajes de datos

Cuadro 6.9: Análisis estadístico de F1-Score por porcentaje de datos

Porcentaje	H-estadístico	$\epsilon_{Comb}^2$	$\epsilon_{Arc}^2$	Índice D	Significancia
1 %	173.84	0.7156	0.0171	41.85	p <0.001
5 %	198.52	0.8225	0.0256	32.13	p <0.001
10 %	208.34	0.8387	0.0234	35.84	p <0.001

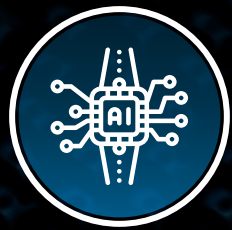
Comparaciones post-hoc por porcentaje

Cuadro 6.10: Comparaciones Mann-Whitney significativas para F1-Score

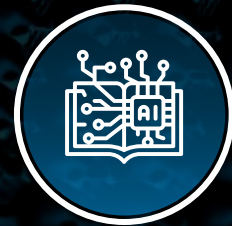
Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	44	66.7 %
5 %	66	41	62.1 %
10 %	66	36	54.5 %



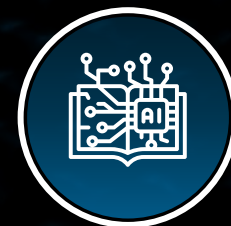
# Conclusiones



**Las características vocales son significativamente más influyentes que la arquitectura neuronal**



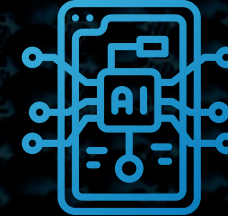
**Es mas eficiente centrar los esfuerzos de investigación en mejorar la extracción de características vocales que complicar las redes neuronales**



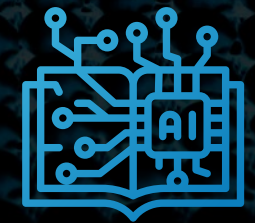
**Los resultados permiten rechazar completamente la hipótesis de igualdad entre grupos**



# Posibles mejoras y Limitaciones



- Ampliar fuentes de datos (idiomas, disfonías)
- Arquitectura especializada



- La tecnología de clonación avanza muy deprisa
- No podemos obtener errores 0 ni combinaciones perfectas



**GRACIAS**