ISSN 1870-4069

Sistema de reconocimiento de voz humana y sintética

Jibran Zaedt Rodriguez Garcia, Andrea Magadán Salazar

TecNM/CENIDET
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET),
México

{m23ce077, andrea.ms}@cenidet.tecnm.mx

Resumen. El advenimiento de la clonación de voz por inteligencia artificial ha revolucionado el campo de la síntesis de voz, ofreciendo una autenticidad y personalización sin precedentes. Las aplicaciones de esta tecnología son numerosas y diversas, abarcando el sector del entretenimiento, la accesibilidad, el marketing digital y enfoques pioneros para la creación de contenido y la comunicación digital. Este artículo analiza algunas de las principales plataformas de clonación de voz y sus aplicaciones prácticas. Examina la funcionalidad, características e impacto de cada plataforma en la comunicación. Luego, la discusión aborda las consideraciones éticas clave que rodean su uso, incluyendo el fraude, la privacidad, la desinformación y el impacto potencial en el empleo de los artistas de voz. El potencial de la clonación de voz para ofrecer experiencias completamente nuevas en muchas áreas es significativo, incluido el uso de voces de personajes en películas y la creación de audiolibros con voces históricas. Además, también proporciona información sobre los motores y algoritmos subyacentes que impulsan estas aplicaciones. También explica cómo la integración de redes neuronales con modelos de alto nivel permite la personalización de voces digitales. Finalmente, el artículo discute la necesidad de un marco ético y regulatorio que garantice prácticas responsables de clonación de voz que protejan los derechos individuales y el valor del trabajo humano en el entorno tecnológico en evolución.

Palabras clave: Herramientas de clonación de voz, algoritmos de clonación, aplicaciones de clonación, inteligencia artificial.

Human and Synthetic Speech Recognition System

Abstract. The advent of artificial intelligence voice cloning has revolutionized the field of speech synthesis, offering unparalleled authenticity and personalization. The applications of this technology are numerous and diverse applications, spanning the entertainment sector, accessibility, digital marketing, and pioneering approaches to content creation and digital communication. This paper analyses some of the main voice cloning platforms and their practical applications. It examines the functionality, features, and impact of

each platform on communication. The discussion then moves on to address the key ethical considerations surrounding their use, including fraud, privacy, misinformation, and the potential impact on the employment of voice artists. The potential for voice cloning to offer entirely new experiences in many areas is significant, including the use of character voices in films and the creation of audio books with historical voices. Furthermore, it also provides insights into the underlying engines and algorithms that power these applications. It also explains how the integration of neural networks with high-level models enables the customization of digital voices. Finally, the paper discusses the need for an ethical and regulatory framework to ensures responsible voice cloning practices that protect individual rights and the value of human labour in the evolving technological environment.

Keywords: Voice cloning tools, cloning algorithms, cloning applications, artificial intelligence.

1. Introducción

En la actualidad, los asistentes de voz o altavoces inteligentes están ganando protagonismo y ya forman parte de nuestra vida. Por ejemplo, se pueden controlar otros dispositivos como termostatos, aires acondicionados, luces, refrigeradores, televisores, entre otros, mediante la voz. Los teléfonos inteligentes han sido los precursores en estos desarrollos de la voz como elemento primordial. Las aplicaciones son variadas y van desde los asistentes personales para ejecutar órdenes, para reconocer dictados, leer textos, para añadir realismo en los videojuegos, etc.

Los sistemas de reconocimiento de voz y generación de voz que pueden procesar de forma natural un diálogo entre humano y máquina se encuentran bajo el término de sistemas o interfaces de voz natural que se destinan a aplicaciones de cliente-servidor en entornos conversacionales. La generación de voz es el proceso mediante el cual un dispositivo inteligente produce secuencias de habla artificial. El reconocimiento de voz es el proceso mediante el cual las computadoras interpretan y digitalizan las señales de voz, tanto para su análisis de contenido como para la interpretación de órdenes.

El proceso de conversión de texto a voz en los asistentes virtuales consta de tres etapas principales [1]:

1. **Entrada de texto y conversión fonémica:** El texto se transforma en una cadena de fonemas, incluyendo puntuación y límites de palabras. Esto permite al modelo capturar mejor la prosodia y los ritmos del habla.
2. **Creación de espectrogramas Mel con Tacotron:** Los fonemas se convierten en espectrogramas Mel utilizando una red basada en la atención secuencia a secuencia, Tacotron. Esta red emplea el enfoque secuencia a secuencia con capas semejantes a la red Long Short-Term Memory en la secuencia para procesar y generar muchos fotogramas del espectrograma al mismo tiempo, lo que lo hace más eficiente y de alta calidad.

3. **Conversión a audios con WaveRNN:** El espectrograma de Mel alimenta a una red neuronal autorregresiva llamada WaveRNN para generar audio, muestra por muestra. En ese nivel, la señal de audio se genera a partir del espectrograma mismo por la red neuronal, y el control de la velocidad y la calidad se realiza mediante una optimización adicional. Este proceso no utiliza los datos de voz exactos que uno podría desear simular. En cambio, utiliza una voz de archivo en el proceso de creación de los fonemas asociados para el texto dado y, por lo tanto, realmente pierde mucho en términos de fidelidad para simular la voz de una persona real.

En los últimos años ha surgido otra área de desarrollo conocida como clonación de voz. En términos sencillos, la clonación de voz es el proceso de copiar la voz de una persona para reproducirla o generarla en un contexto diferente al original. Es producir una voz artificial que tenga las mismas características (suene igual) como si la hubiera pronunciado una persona objetivo [1].

La clonación de voz no es nueva; sin embargo, las nuevas herramientas de inteligencia artificial logran mayores niveles de autenticidad y personalización. Este progreso se ha utilizado para replicar la voz humana de formas que ensalzan en sectores como el entretenimiento, la accesibilidad y el marketing digital. La clonación de voz está liderando actualmente la creación de contenido y las experiencias de comunicación digital [2]. Si bien algunas aplicaciones consideran texto para la clonación de voz, no se recomienda porque suele provocar pérdida de información en el proceso de transmisión oral de un mensaje.

El objetivo de este artículo es presentar las principales plataformas para la clonación de voz y sus aplicaciones en la vida real, así como analizar el uso de la frecuencia fundamental (f_0) como característica principal en el entrenamiento de un clasificador basado en Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) para la identificación de voces clonadas y naturales. Se exploran las capacidades de esta metodología en la detección de patrones distintivos entre voces generadas artificialmente y voces humanas reales, con el fin de evaluar su eficacia en la autenticación y verificación de identidad en entornos digitales.

Los sistemas revisados se basan en las tendencias de las comunidades activas como: Discord [3], GitHub [4] y similares, donde los desarrolladores y entusiastas comparten su experiencia de primera mano y sus preferencias por dichas técnicas. Se pretende examinar cómo funcionan las herramientas, en qué entornos y qué cambian en la comunicación digital. Esto hace que la revisión de las herramientas de clonación de voz sea relevante no solo para investigadores y desarrolladores, sino también para cualquier persona que esté interesada en los avances en la Inteligencia Artificial y sus consecuencias para las sociedades.

Para comprender mejor el potencial de la clonación de voz, a continuación, se presenta una lista de aplicaciones útiles de esta tecnología en diversos sectores.

2. Herramientas de clonación de voz

A diferencia de los avances en la detección de deepfakes en imágenes, se observa una escasez de trabajos dedicados a la detección de voces clonadas [5]. Esto subraya

la necesidad de investigaciones adicionales en este campo para abordar los desafíos específicos asociados con la manipulación de audio.

La importancia de este campo radica en la necesidad de proporcionar al usuario final una mayor confianza en los sistemas de comunicación y verificación de identidad. A medida que las técnicas de falsificación de voces se vuelven más sofisticadas, se incrementa también la urgencia de desarrollar métodos de detección igualmente avanzados. Esto no solo contribuye a la seguridad personal y empresarial, sino que también juega un papel crucial en la preservación de la integridad de la información y la prevención del fraude.

Se llevó a cabo una investigación sobre varias herramientas, de preferencia gratuitas, que pueden generar audios de voces clonadas de manera convincente, evitando sonidos ”antinaturales”. Las siguientes herramientas de clonación de voz tienen distintas capacidades que se utilizan según las necesidades del usuario y/o de las aplicaciones:

1. **ElevenLabs** [6]: Es una herramienta avanzada de Inteligencia Artificial (IA) que ofrece tecnologías de Texto a Voz, Voz a Voz y Clonación de voz. Con esta aplicación es posible generar audio hablado de alta calidad en una variedad de voces, estilos e idiomas (actualmente 32). También permite ajustar géneros, edades, tonos y acentos según las preferencias del usuario.

   Su modelo de IA captura de manera excepcional la entonación y las inflexiones humanas, ofreciendo una experiencia de voz sumamente realista. Para evitar el uso de su tecnología en la creación de deepfakes, ElevenLabs ha adoptado controles, permitiendo que este producto esté disponible solo para usuarios verificados mediante suscripción.

   Utiliza algoritmos que maximizan la estabilidad y similitud de las voces, ajustables a través de su API. ElevenLabs ha proporcionado en GitHub la documentación y los ejemplos de código necesarios para integrarlo con herramientas como Python y Java. Los modelos disponibles en esta plataforma incluyen Multilingual v2, English v1, Turbo v2 y Turbo v2.5.

2. **VocaliD** [7]: Crea voces personalizadas para personas con discapacidades del habla. Para la generación de voz combina las características vocales de los usuarios con voces pregrabadas para generar una voz única. Esto lo logra mediante la combinación de su base de datos ”Human Voicebank”, que incluye más de 14.000 donantes en 110 países.

   La integración de estas voces en dispositivos de asistencia como Tobii Dynavox dice mucho sobre la personalidad y las emociones del usuario.

3. **Applio** [8]: Es una aplicación de clonación de voz de uso gratuito sin límites para crear su modelo de clonación de voz. Permite la síntesis de voz a texto y de voz a voz. Realiza la transformación de audio utilizando diferentes algoritmos de extracción de tono como Pitch Marking, Harvest, DIO, Rmvpe y Rmvpe_gpu.

   Permite varias opciones para ajustar el procesamiento de audio, lo cual brinda variedad entre ser una herramienta experimental o una aplicación más profesional para fines específicos. Esta aplicación es de uso público y gratuito, mediante la cual los usuarios pueden crear modelos de clonación de voz sin límite alguno. Además,

ofrece otras herramientas como la descarga de modelos de voces ya entrenados y listos para ser utilizados.

4. **RVC** [9]: Esta aplicación cuenta con una versión web y está disponible para el público en general de forma gratuita. A través de ella, los usuarios pueden crear modelos de clonación de voz sin restricciones.

   Además de esta función principal, la aplicación ofrece diversas opciones adicionales como cambiar el audio de la grabación seleccionando la frecuencia de muestreo (40k o 48k). También permite seleccionar el algoritmo de extracción, que puede ser Pitch Marking, Harvest, DIO o Rmvpe, con opciones de 0 a 8 subprocesos de CPU. Es una alternativa flexible para quienes necesitan personalizar los modelos de voz.

5. **Voice.ia** [10]: Esta aplicación cuenta con una amplia variedad de voces desarrolladas por la comunidad. Puede proporcionar un cambio de voz en tiempo real y da acceso a una gran cantidad de voces creadas y almacenadas.

   Es una aplicación paga, aunque incluye la opción de recibir una paga mínima diaria por iniciar sesión, que se puede usar para comprar más voces. Las voces se pueden usar sin restricciones una vez compradas. No se menciona ningún algoritmo de extracción, lo que puede afectar a los usuarios que buscan detalles técnicos específicos.

3. Algoritmos utilizados por las herramientas

Los algoritmos forman una parte esencial en la clonación de voces porque son la base de la construcción de las herramientas en el análisis, ajuste y reproducción de las voces con precisión. Las técnicas aplicadas incluyen:

– **Redes neuronales profundas**: Se utilizan para analizar y reproducir características vocales.

– **Pitch Marking**: El algoritmo se utiliza en el procesamiento del habla para detectar cambios en la frecuencia fundamental o el tono. Es útil para analizar la calidad de la voz y en la síntesis de voz.

– **Harvest**: Un algoritmo de extracción de tono que se utiliza para aplicar el tono de la voz original a la voz clonada.

– **DIO**: Es un método para estimar la frecuencia fundamental, son técnicas de procesamiento paralelo o distribuido en sistemas informáticos.

– **Modelo robusto para la estimación del tono vocal en música polifónica**: Se utiliza para estimar el tono vocal en música polifónica.

– **Modelos multilingües**: Estos modelos incluyen Multilingual v1 y Multilingual v2, que ofrecen estabilidad y soporte para 29 idiomas.

– **Turbo**: Turbo v2 y Turbo v2.5 son algoritmos de baja latencia optimizados para conversaciones en tiempo real, diseñados para quienes requieren hablar de manera rápida y sencilla.

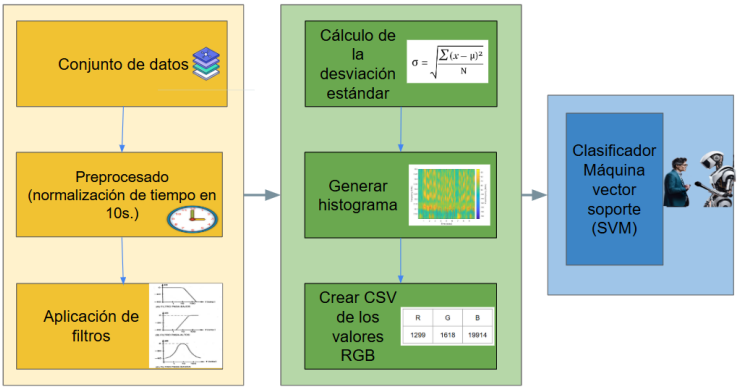

Fig. 1. Metodología propuesta.

– **CREPE**: Es un algoritmo de seguimiento de tono monofónico basado en una red neuronal simple para lograr la segmentación de notas monofónicas.

Conocer los algoritmos básicos detrás de la clonación de voz permite apreciar en profundidad cómo estas herramientas logran resultados precisos y de alta calidad.

3.1. Metodología

Como se puede apreciar en la figura 1, la metodología propuesta consta de tres etapas fundamentales:

1. **Procesamiento del conjunto de datos:**
   – Todos los audios fueron normalizados a una duración máxima de 10 segundos para garantizar uniformidad en el procesamiento.
   – Se verificó que todos los archivos estuvieran en formato .WAV; los que no cumplían con esta condición fueron convertidos utilizando scripts en Python.
   – Se aplicaron filtros digitales de tipo Butterworth de orden 5 con los siguientes parámetros:
     • Filtro pasa baja: Atenuación de frecuencias superiores a 500 Hz.
     • Filtro pasa alta: Atenuación de frecuencias inferiores a 2000 Hz.
     • Filtro pasa banda: Permite el paso de frecuencias entre 500 Hz y 2000 Hz.

     Estos filtros permiten resaltar componentes relevantes de la voz humana y reducir el ruido.
2. **Extracción de características:**
   – Se extrajo la frecuencia fundamental (F0) de cada audio utilizando un modelo convolucional especializado (convModel.mat), previamente entrenado para la estimación de tono en señales monofónicas [11].

- A partir de los valores de F0 extraídos, se construyeron histogramas que registran la distribución de frecuencias. Estos histogramas fueron representados mediante los canales de color Rojo (R), Verde (G) y Azul (B), lo cual permite organizar los datos en una estructura matricial.
- Como se menciona en el artículo [12], los histogramas de características acústicas pueden ser utilizados para el análisis de señales de voz. Sin embargo, a diferencia de dicho enfoque, que aplica una transformada de Fourier y una red neuronal para procesar la imagen del histograma, en el presente trabajo se utilizan directamente los valores de intensidad de píxeles en los canales RGB como vectores de entrada para el clasificador Máquina de Vectores de soporte (VSM por sus siglas en inglés), permitiendo una clasificación binaria entre voces naturales y clonadas sin recurrir a arquitecturas profundas.
- Esta estrategia permite capturar variaciones relevantes en la distribución de F0 de manera estructurada, facilitando al clasificador la detección de patrones distintivos entre ambas clases de voz.

3. **Clasificación:**
   - Los histogramas procesados fueron utilizados como vectores de características para entrenar un clasificador VSM.
   - Se probaron diferentes núcleos: lineal, polinomial, radial (RBF) y sigmoidal.
   - Se utilizó el 80 % del dataset para entrenamiento y el 20 % para validación.
   - La métrica principal de evaluación fue la exactitud, alcanzando un 95.54 % de exactitud utilizando filtro pasa baja y kernel sigmoidal.

4. Conjunto de datos

4.1. Conjunto de datos de voces naturales

CommonVoice [13] es un proyecto desarrollado por Mozilla que tiene como objetivo crear un conjunto de datos de voz abierto y diverso, destinado a mejorar la accesibilidad y la representación en las tecnologías de reconocimiento de voz. Este recurso se construye mediante las contribuciones de voluntarios, quienes participan grabando y validando frases en distintos idiomas.

Principales características:

- **Datos abiertos:** Las grabaciones recopiladas se distribuyen bajo la licencia CC0 (dominio público), lo que garantiza su disponibilidad sin restricciones. Esto permite que los datos sean utilizados en investigación, desarrollo de software y en la implementación de tecnologías de voz.
- **Idiomas y diversidad:** El proyecto incluye soporte para más de 100 idiomas y está diseñado para capturar una amplia gama de acentos y dialectos, lo que contribuye a reflejar la diversidad lingüística y cultural a nivel global.

Tabla 1. Conjunto de datos de voces clonadas.

Conjunto	Tiempo de audios	Número de audios
Voces Clonadas Español	149 minutos	796
Voces Clonadas Inglés	97 minutos	460

Tabla 2. RVC Dataset.

Idioma	Género	Cantidad
Español	Mujer	100
	Hombre	100
	Subtotal	200
Inglés	Mujer	60
	Hombre	100
	Subtotal	160
Total		360

Tabla 3. Applio Dataset.

Idioma	Género	Cantidad
Español	Mujer	56
	Hombre	100
	Subtotal	156
Inglés	Mujer	64
	Hombre	140
	Subtotal	204
Total		360

– **Contribuciones voluntarias:** La plataforma permite que cualquier persona participe leyendo frases para grabar su voz o revisando las grabaciones de otros para validar su calidad. Este enfoque colaborativo es fundamental para la construcción del conjunto de datos.

– **Propósito del proyecto:** La iniciativa busca democratizar el acceso a las tecnologías de voz, reduciendo sesgos en los sistemas existentes y promoviendo herramientas inclusivas que representen a una mayor variedad de usuarios.

4.2. Conjuntos de datos de voces generados artificialmente

Se plantea la necesidad de abordar la limitación encontrada en los conjuntos de datos existentes de voces clonadas, los cuales, como se mencionó anteriormente, no son óptimos, ya que son distinguibles [14]. Por consiguiente, se realizó la generación de un nuevo conjunto de datos que abarca aproximadamente 1300 audios clonados, como se muestra en la tabla 1. Para este propósito, se utilizan 14 audios por voz, lo que resulta en un estimado de 100 voces diferentes, tanto masculinas como femeninas.

Los audios presentes en este dataset son de tiempos variables, desde los 11 segundos hasta los 2 minutos. En las tablas 2, 3, 4, y 5 se muestra información de los conjuntos de datos.

Tabla 4. Conjunto de datos Elevenlabs.

Idioma	Género	Cantidad
Español	Mujer	60
	Hombre	40
	Subtotal	100
Inglés	Mujer	40
	Hombre	60
	Subtotal	100
Total		200

Tabla 5. Conjunto de datos Voice.ia.

Idioma	Género	Cantidad
Español	Mujer	40
	Hombre	60
	Subtotal	100
Inglés	Mujer	200
	Hombre	36
	Subtotal	236
Total		336

4.3. Procesado de los audios

Se verifica que los archivos de audio estén en formato ".WAV", por lo que el primer paso consiste en transformar aquellos que no cumplen con este requisito al formato adecuado. Para ello, se emplea un código en Python que analiza todos los archivos de audio en la carpeta y, en caso de ser necesario, realiza la conversión correspondiente.

A continuación, se aplican los filtros de "pasa baja", "pasa alta" y "pasa banda" [15] a los audios en formato correcto. Posteriormente, se calcula la frecuencia fundamental y se generan los histogramas, los cuales se almacenan en un archivo de texto.

Los datos de los histogramas se dividen en los canales "R", "G" y "B", registrando la frecuencia de los valores en un rango de 0 a 255 dentro de un archivo de Excel. Final- mente, estos valores son utilizados en una máquina de soporte vectorial, creando un conjunto de datos para cada filtro por separado.

5. Etapa 2: Extracción de características

En [16] se señala que existe una diferencia significativa entre la frecuencia fundamental (F0) promedio de hombres y mujeres. En términos generales, la F0 promedio es notablemente más alta en mujeres y el rango de F0 en Hz es más amplio en comparación con el de los hombres. A pesar de que las diferencias en el rango de F0 entre sexos desaparecen cuando se expresan en semitonos o como un factor de modulación, los valores promedio de F0 continúan siendo distintos. En resumen, la F0 promedio de los hombres no es equivalente a la de las mujeres.

El factor de modulación, en el contexto de la voz, se refiere a la variación de la frecuencia fundamental (F0) de la voz a lo largo de un período de tiempo. Esta medida indica cuánto fluctúa la frecuencia fundamental alrededor de su valor

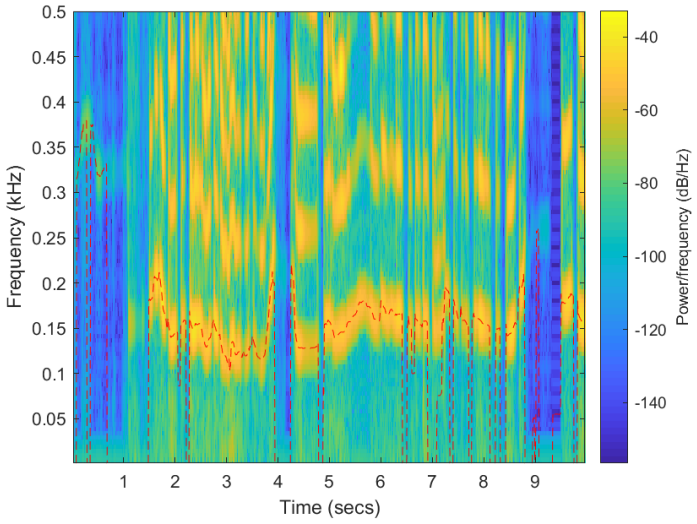

Fig. 2. Histograma de voz natural.

promedio. La modulación de frecuencia puede ser cuantificada en porcentaje o en unidades de frecuencia (como Hz). Un alto factor de modulación sugiere que la voz presenta numerosas variaciones en tono, lo cual puede interpretarse como expresividad o variabilidad vocal.

Teniendo en cuenta lo anterior, se usó el modelo de `Crepe-tiny` [11] para el procesamiento de $hf0$, el cual es un rastreador de tono monofónico basado en una red neuronal convolucional poco profunda que opera sobre la función de autocorrelación normalizada en el dominio del tiempo. Dicho código ha sido modificado para ser capaz de procesar diversos datos, ya que su versión base trabajaba únicamente con un audio a la vez y no se guardaban los histogramas generados ni el vector de datos.

En la figura 2 se puede observar cómo el histograma muestra en el eje X el tiempo en segundos, desde los 0 hasta los 10, en color se muestra la potencia y en el eje Y se muestra la frecuencia de la voz.

6. Experimentación

6.1. Especificaciones de hardware, versión de Python y librerías utilizadas

El equipo utilizado para desarrollar el proyecto consta de un CPU Intel i7 -600K, 32Gb de memoria RAM DDR4 y una GPU NVIDIA GeForce RTX 3080 Ti de 12Gb de VRAM. Se trabajó con Python 3.9.13, las librerías utilizadas junto a su versión y una descripción corta se encuentran en la tabla 6.

Se seleccionó el uso de Máquinas de Vectores de Soporte (SVM) debido a su alta capacidad para resolver problemas de clasificación binaria en espacios de alta dimensión, como el caso de la representación mediante histogramas de F0 [17], con los kernels “lineal, polinomial, gaussiano y sigmoide”, haciendo uso de los hiper

Tabla 6. Librerías utilizadas en el proyecto.

Librería	Versión	Descripción
matplotlib	3.9.2	Librería de visualización para crear gráficos en 2D.
numpy	1.23.5	Librería para cálculos numéricos.
pandas	1.5.3	Librería para análisis y manipulación de datos.
librosa	0.10.2.post1	Librería para el procesamiento de audios.
CUDA	11.2.67	Librería para hacer uso de GPU.
crepe	0.0.16	Librería para calcular f0 de los audios.
sklearn	1.6.1	Librería para el uso de máquina de soporte vectorial.

Tabla 7. Resultados con la exactitud.

Características	Kernel	Exactitud
Pasa banda	Lineal	0.9018
Pasa banda	Polinomial	0.8661
Pasa banda	RBF	0.9018
Pasa banda	Sigmoideo	0.8304
Pasa alto	Lineal	0.8571
Pasa alto	Polinomial	0.8750
Pasa alto	RBF	0.8571
Pasa alto	Sigmoideo	0.8393
Pasa baja	Lineal	0.9018
Pasa baja	Polinomial	0.8482
Pasa baja	RBF	0.8571
Pasa baja	Sigmoideo	0.9554
Sin filtro	Lineal	0.9018
Sin filtro	Polinomial	0.8839
Sin filtro	RBF	0.9018
Sin filtro	Sigmoideo	0.9196

parámetros por defecto en la librería sklearn. Del conjunto de datos se utilizó el 80 % para entrenamiento y el 20 % para validación. En la tabla 7, se observan los resultados de los filtros con todos los kernels utilizados.

El análisis de los resultados obtenidos mediante la aplicación de diferentes filtros y kernels de SVM revela patrones significativos que pueden guiar la selección del método más adecuado para tareas específicas de clasificación de señales.

– **Filtro Pasa Banda:** El filtro pasa banda muestra un rendimiento consistente con los kernels lineal y RBF, ambos alcanzando una exactitud de 0.9018. Sin embargo, el kernel RBF logra una sensibilidad perfecta (1.0000), lo que sugiere una capacidad superior para identificar correctamente las señales positivas, aunque con una especificidad moderada. Este kernel podría ser preferido en aplicaciones donde es crucial minimizar los falsos negativos.

– **Filtro Pasa Alta:** En el caso del filtro pasa alta, el kernel polinomial presenta un buen equilibrio entre exactitud (0.8750) y alta sensibilidad (0.9286), pero con una especificidad constante de 0.8214. Esto indica que el kernel polinomial es eficiente en la detección de verdaderos positivos, aunque su capacidad para discriminar los

verdaderos negativos es limitada. Este filtro y kernel pueden ser útiles en escenarios donde la detección correcta de señales es prioritaria sobre la especificidad.

– **Filtro Pasa Baja:** El filtro pasa baja combinado con el kernel sigmoide destaca al lograr la mayor exactitud (0.9554) y alta especificidad (0.9286), lo que sugiere una excelente capacidad tanto para identificar correctamente las señales positivas como para minimizar los falsos positivos. Este método sería ideal para aplicaciones que requieren una alta exactitud general y una excelente discriminación entre señales positivas y negativas.
– **Sin Filtro:** Sin la aplicación de un filtro, el kernel sigmoide también muestra un rendimiento notable, con una exactitud de 0.9196 y una alta especificidad (0.9464). Esto sugiere que este método es altamente efectivo para aplicaciones donde se necesita un análisis detallado de las características de frecuencia de las señales.

7. Conclusión

Los resultados obtenidos en este estudio demuestran que las características extraídas del histograma de la frecuencia fundamental ($f0$) son una herramienta eficaz para diferenciar entre voces clonadas y naturales. El uso de un clasificador basado en Máquinas de Vectores de Soporte (SVM) permitió alcanzar una precisión del 95 %, lo que indica un alto nivel de acierto en la identificación de voces generadas artificialmente. Estos hallazgos validan la utilidad de la frecuencia fundamental como una característica distintiva en el análisis de señales de voz y refuerzan su potencial en aplicaciones de autenticación y detección de fraudes en entornos digitales. Con estos resultados, se concluye que el objetivo del artículo se ha cumplido satisfactoriamente, destacando la relevancia de esta metodología en el estudio y la seguridad de los sistemas de clonación de voz.

Actualmente, el conjunto de datos empleado no ha sido publicado debido a que forma parte de un trabajo de investigación en curso dentro de un programa de maestría. Aún se están evaluando distintas metodologías y realizando experimentaciones complementarias. Una vez concluido el proceso académico y obtenido el grado correspondiente, se planea poner a disposición de la comunidad científica el dataset completo junto con los scripts utilizados, con el fin de favorecer la reproducibilidad y futuras investigaciones en este campo.

Agradecimientos. Se agradece a SECIHTI por el apoyo económico brindado mediante la beca para los estudios de maestría.

Referencias

1. S. Achanta., R. Maas., R. Clark.: On-Device Neural Speech Synthesis. arXiv, pp. 1–7 (2021), doi: 10.48550/arXiv.2109.08710.
2. Extracta.ai: Exploring the Impact of AI Voice Cloning: Transforming Digital Storytelling. URL: https://extracta.ai/exploring-the-impact-of-ai-voice-cloning-transforming-digital-storytelling (2024)

3. Discord: Discord. URL: https://discord.com (2022)
4. GitHub: GitHub. URL: https://github.com (2024)
5. Meta AI: Deepfake Detection Challenge Results: An Open Initiative to Advance AI. URL: https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai (2023)
6. Eleven Labs: Eleven Labs. URL: https://elevenlabs.io (2024)
7. VocaliD: Vocalid. URL: https://vocalid.ai (2024)
8. GitHub: Iahispano/applio-rvc-fork. URL: https://github.com/IAHispano/Applio-RVC-Fork (2023)
9. GitHub: RVC-Project/Retrieval-based-Voice-Conversion-WebUI. URL: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI (2023)
10. Voice.ai: Voiceai. https://voice.ai (2023)
11. Pradeepiit: Hf0. URL: https://github.com/Pradeepiit/hf0 (2024)
12. Lim, S.-Y., Chae, D.-K., Lee, S.-C.: Detecting Deepfake Voice Using Explainable Deep Learning Techniques. Applied Sciences, 12(8), pp. 1–15 (2022) doi: 10.3390/app12083926.
13. Mozilla: Common Voice. URL: https://commonvoice.mozilla.org/en/datasets (2024)
14. Blue, L., Warren, K., Abdullah, H.: Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In: 31st USENIX Security Symposium, pp. 2691–2708 (2022) URL: https://www.usenix.org/conference/usenixsecurity22/presentation/blue
15. Corchete, V.: High-Pass, Low-Pass and Band-Pass Filtering. Universidad de Almería, pp.1–5 (2019) doi: 10.13140/RG.2.2.25817.67686.
16. Traunmüller, H., Eriksson, A.: The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults. URL: https://www.researchgate.net/publication/240312210 (1995)
17. Carmona, E. J.: Tutorial sobre Máquinas de Vectores Soporte (SVM). pp. 1–27. URL: https://www.researchgate.net/publication/263817587.