# Biometría de Voz: Características de la Voz Humana y Técnicas Anti‑Spoofing

## Introducción

La biometría de voz es la tecnología que permite verificar o identificar a una persona mediante las características únicas de su voz. A diferencia del **reconocimiento de voz** (speech recognition) que interpreta *qué* se dice, la biometría de voz se centra en *quién* lo dice[[1]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=concentrates%20on%20choosing%20the%20proper,can%20be%20identified%20or%20verified). El **señal de la voz humana** contiene abundante información distintiva: no solo la identidad del hablante, sino también su idioma nativo, estado físico o emocional[[2]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=Automatic%20speaker%20recognition%20is%20an,physical%20state%2C%20and%20emotional%20state). Estas señales distintivas se extraen únicamente del **audio capturado por un micrófono**, sin necesidad de mediciones fisiológicas directas. Un sistema biométrico de voz analiza dichos rasgos acústicos para crear una “huella vocal” única de cada individuo.

Un desafío importante es que la voz de una persona puede variar por factores temporales o de salud (por ejemplo, una **gripa o resfriado**). Incluso en condiciones normales, la voz presenta variaciones día a día[[3]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=%60%60Cold,from%20each%20person%20to%20train). Durante un resfriado común, suelen aparecer diferencias audibles: la voz puede sonar más nasal, ronca o con tos, y puede haber ligeras variaciones en el tono (pitch)[[4]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=the%20normal%2C%20healthy%20signals,the%20acoustic%20signal%20and%20the). Un sistema robusto debe aprovechar características vocales profundas que permanecen relativamente estables, de modo que pueda **reconocer al hablante incluso si está enfermo**. (Cabe mencionar que en casos extremos, como una laringitis severa, la alteración vocal puede ser tan grande que dificulte la coincidencia de la huella vocal[[5]](https://www.iproov.com/es/blog/disadvantages-vulnerabilities-voice-biometrics#:~:text=dificultando%20o%20imposibilitando%20a%20algunas,hacer%20coincidir%20la%20huella%20vocal), pero una simple gripa típicamente no impide la identificación si el sistema está bien entrenado).

A continuación, detallaremos **las principales características del audio de la voz humana** empleadas para identificar de forma única a un hablante, seguidas de las **técnicas de anti-spoofing** (anti-suplantación) por software diseñadas para prevenir fraudes mediante voces grabadas o sintéticas.

## Características de la voz humana para identificación biométrica

El sonido de la voz de cada persona posee rasgos acústicos únicos que sirven como identificadores biométricos. Estas características se pueden extraer de la señal de audio y actúan como parámetros de comparación entre la voz de un usuario y las huellas vocales almacenadas en la base de datos. A continuación se listan las **características propias del audio de la voz** más importantes, junto con sus definiciones y su relevancia:

* **Frecuencia fundamental (tono o *pitch*):** Es la frecuencia base de vibración de las cuerdas vocales y se percibe como la altura tonal de la voz. Corresponde al **tono hablado** característico de cada persona. Suele ser un valor promedio (por ejemplo, voces masculinas tienden a F0 más baja que voces femeninas), pero cada individuo tiene un rango tonal propio. El tono es importante para identificar la voz, aunque pueda variar ligeramente con emoción o enfermedad[[6]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,resonantes%20que%20caracterizan%20el%20timbre). Incluso con un resfriado, la frecuencia fundamental de un hablante generalmente se mantiene dentro de un rango reconocible, salvo casos de afonía grave.
* **Formantes (resonancias vocales):** Son las **frecuencias resonantes** del tracto vocal que dan forma al espectro sonoro de la voz. Se designan típicamente como F1, F2, F3, F4, etc. Los formantes dependen de la anatomía y la configuración de la boca, nariz y garganta del hablante, y *caracterizan el timbre o calidad tonal* de su voz[[7]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,Root%20Mean). En otras palabras, son picos en el espectro de frecuencias que distinguen a cada voz (y también distinguen vocales en el habla). Los valores específicos de formantes de una persona tienden a ser únicos y relativamente estables; aunque una gripa puede introducir cierta nasalidad (afectando ligeramente algunos formantes), el patrón general de resonancias permanece suficientemente distintivo para la identificación.
* **Timbre (calidad tonal):** Es la **cualidad sonora** que nos permite distinguir dos voces aunque pronuncien la misma nota. El timbre está determinado por el contenido armónico y las características espectrales de la voz (principalmente los formantes y la distribución de energía en el espectro). En resumen, refleja la “firma acústica” global del hablante. Por ejemplo, voces aterciopeladas, nasales, ásperas, etc., se diferencian por su timbre. Esta propiedad es capturada matemáticamente mediante la envolvente espectral de la señal[[8]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=%2F%20Segundo%20semestre%202024%2056,permite%20capturar%20la%20esencia%20de). El timbre suele mantenerse identificable incluso si la persona está congestionada; aunque una congestión nasal puede hacerlo más opaco, muchos rasgos tonales permanecen constantes.
* **Coeficientes Cepstrales de Frecuencia Mel (MFCC):** Son un conjunto de parámetros ampliamente utilizados que **representan la envolvente del espectro de la señal de voz** de forma compacta[[8]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=%2F%20Segundo%20semestre%202024%2056,permite%20capturar%20la%20esencia%20de). Los MFCC capturan el timbre y resonancias fundamentales percibidas por el oído humano, descartando información redundante. En biometría de voz, los MFCC son el estándar de facto para caracterizar la voz, pues concentran las características acústicas distintivas en ~12–20 coeficientes que alimentan los algoritmos de reconocimiento[[9]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=En%202019%2C%20se%20desarroll%C3%B3%20una,1). Estos coeficientes, al codificar el patrón espectral, suelen ser robustos a variaciones menores (por ejemplo, un resfriado podría afectar ligeramente la voz, pero la “forma” general del espectro en los MFCC seguirá correspondiendo al mismo hablante). De hecho, numerosos sistemas de identificación text-independent se basan en MFCC como entrada, complementados a veces con otros coeficientes como LPC, PLP, etc., para mejorar la robustez[[9]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=En%202019%2C%20se%20desarroll%C3%B3%20una,1).
* **Energía e intensidad de la voz:** La **energía** de la señal mide la potencia o volumen global de la voz. La **intensidad**, a menudo expresada en términos de nivel RMS, se refiere al nivel de amplitud promedio de la voz[[10]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,dispersi%C3%B3n%20de%20las%20frecuencias%20en). Algunas personas hablan naturalmente más fuerte o más suave que otras, lo cual puede formar parte de su perfil de voz. Sin embargo, la intensidad puede variar con la distancia al micrófono o el estado de ánimo, por lo que por sí sola no es un identificador confiable. Aun así, se suele incluir la energía como uno de los parámetros acústicos a analizar[[11]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identi,Tono%20%28%20pitch), y en combinación con otros rasgos puede ayudar (por ejemplo, midiendo la distribución de energía en distintas bandas de frecuencia). Un sistema robusto normaliza en lo posible las diferencias de volumen, para enfocarse en rasgos más intrínsecos.
* **Entropía espectral:** Es una medida de la **complejidad y variabilidad del espectro** de la voz[[11]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identi,Tono%20%28%20pitch). Indica cuán uniforme o dispersa está la energía a través de las frecuencias. Cada hablante puede tener un espectro más “plano” o más “pico” dependiendo de la riqueza armónica y articulación de su voz. Esta característica puede contribuir a diferenciar voces (por ejemplo, una voz muy modulada con muchos armónicos vs. una voz más pura con pocos picos espectrales). La entropía espectral promedio se usa como descriptor cuantitativo de la **calidad de la voz**.
* **Variabilidad de frecuencia (rango vocal):** Mide la **dispersión o amplitud del rango de frecuencias** que utiliza un hablante[[12]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Formantes%20%28F1%2C%20F2%2C%20F3%2C%20F4%29,frecuencia%20a%20lo%20largo%20del). En esencia, refleja cuánto varía el tono fundamental durante el habla; algunos individuos mantienen un tono muy constante, mientras otros varían ampliamente (voz más expresiva). También puede relacionarse con los cambios en formantes al articular distintos fonemas. Esta variabilidad está vinculada a la **prosodia** y entonación del habla. Aunque una persona tenga gripa, su tendencia a usar un rango tonal amplio o estrecho normalmente se conserva (salvo que la afección limite severamente su voz).
* **Prosodia (entonación y ritmo):** La prosodia abarca los *patrones melódicos y rítmicos* del habla de una persona. Incluye el **contorno del tono** a lo largo de una frase (subidas y bajadas de pitch), la duración de sonidos y pausas, el énfasis en ciertas sílabas, y la velocidad de habla. Cada hablante desarrolla un estilo prosódico propio: por ejemplo, hay quien habla muy pausado y otro muy atropellado; algunos suben el tono al final de las frases, otros no. La prosodia captura la **variabilidad de la frecuencia a lo largo del tiempo** y el patrón temporal del habla[[13]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Square%20%29,variabilidad%20en%20el%20periodo%20y)[[14]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,del%20tono%20durante%20la%20conversaci%C3%B3n). Características como la **duración promedio de los silencios** (pausas) entre palabras también entran aquí[[11]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identi,Tono%20%28%20pitch). Estos rasgos de *comportamiento del habla* son útiles para reconocimiento de hablante, sobre todo en grabaciones más largas (en pocas palabras puede no apreciarse). Cabe destacar que una condición como la gripa suele tener poco efecto en la prosodia de la persona (uno seguirá tendiendo a pausar o entonar de forma similar); por ello, la prosodia brinda información complementaria para la identificación.
* **Coarticulación:** Es un fenómeno lingüístico donde los sonidos no se pronuncian de forma totalmente independiente, sino solapada; es decir, **cada fonema se ve influenciado por los sonidos anteriores y siguientes** en la cadena hablada[[15]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,partir%20de%20la%20descomposici%C3%B3n%20arm%C3%B3nica). El *patrón de coarticulación* varía sutilmente entre hablantes según su acento y hábitos de pronunciación. Por ejemplo, la manera en que una persona enlaza ciertas consonantes y vocales (más fluido, o marcando separaciones) puede diferir y servir como característica identificativa. En biometría de voz, la coarticulación se considera en sistemas text-independent: al modelar distribuciones de fonemas y transiciones, se capturan esas diferencias idiosincráticas. Aunque la coarticulación puede cambiar con distintos idiomas o si la persona deliberadamente modifica su habla, bajo condiciones normales (incluso con resfriado) cada hablante mantiene sus patrones típicos de encadenamiento de sonidos.
* **Jitter y Shimmer:** Son medidas de la **estabilidad micróscopica de la voz**. El *jitter* cuantifica la variación ciclo a ciclo en la frecuencia fundamental (es decir, inestabilidades en el periodo de vibración de las cuerdas vocales), mientras que el *shimmer* cuantifica la variación en la amplitud de esos ciclos[[16]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Square%20%29,Miden%20caracter%C3%ADsticas%20espec%C3%ADficas%20de). En términos simples, indican cuán “temblorosa” o “estable” es la voz de una persona en tono y volumen, respectivamente. Cada individuo tiene cierto grado natural de jitter/shimmer; por ejemplo, voces muy jóvenes o entrenadas tienden a ser más estables (bajo jitter), mientras que personas mayores o con afecciones pueden mostrar más variaciones. Estas medidas se emplean comúnmente en estudios de patología vocal, pero también pueden integrarse en una huella vocal biométrica para aportar información sobre la **calidad vocal** intrínseca del hablante. Un resfriado moderado podría incrementar ligeramente el jitter (por irritación de la laringe) o el shimmer (por voz más entrecortada), pero la *línea base* de estos valores en cada persona permanece característica en ausencia de cambios drásticos en la voz.
* **Relación Armónico-Ruido, Voceo y Soplo:** La voz humana combina componentes **harmónicos** (producidos por vibración periódica de las cuerdas vocales, percibidos como tono definido) y **ruidosos** (aspiraciones, susurros, fricación, etc.). La **relación armónico/ruido (HNR)** mide la proporción de energía harmónica vs. ruido en la voz. De modo parecido, métricas denominadas *“voceo”* y *“soplo”* se refieren a características de **voz sonorizada vs. aire soplado o aliento** en la emisión[[16]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Square%20%29,Miden%20caracter%C3%ADsticas%20espec%C3%ADficas%20de). Una voz muy limpia y resonante tendrá alto componente armónico (mucho voceo), mientras que una voz susurrante o ronca tendrá más componente de ruido (soplo). Cada persona tiene un timbre con cierto grado de aire al hablar – por ejemplo, hay quienes tienen una voz naturalmente susurrada o con escape de aire audible, frente a quienes tienen una voz muy clara y completamente sonora. Estas proporciones pueden servir para distinguir identidades. Un sistema biométrico podría registrar que la voz de *X* tiene cierto HNR o cierta cualidad aspirada única. Aunque una gripa podría aumentar transitoriamente el componente de ruido (por ejemplo, por disfonía o mucosidad que introducen soplo), la tendencia general de la voz del individuo (más aireada vs más metálica) tiende a conservarse a largo plazo.

En la práctica, los sistemas modernos de reconocimiento de hablante utilizan métodos estadísticos y de aprendizaje profundo que combinan muchas de estas características simultáneamente. Por ejemplo, los algoritmos de **i-vectores** o **x-vectors** generan una representación de alta dimensión de la voz del hablante, encapsulando información de timbre, formantes, prosodia, etc., en un vector único. Sin embargo, todas estas técnicas se basan en los principios fundamentales antes mencionados. Una huella vocal eficaz debe resaltar los rasgos **invariantes** de la persona (los que no cambian fácilmente con la frase pronunciada, el ruido de fondo o una leve enfermedad), maximizando las diferencias entre hablantes[[1]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=concentrates%20on%20choosing%20the%20proper,can%20be%20identified%20or%20verified). Estudios han demostrado que incorporando suficiente variedad de muestras (diferentes días, condiciones, incluso muestras de voz “con gripe”), el modelo puede aprender a reconocer al individuo en **días tanto sanos como enfermos**[[17]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=The%20goal%20of%20this%20research,physical%20state%2C%20and%20emotional%20state)[[3]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=%60%60Cold,from%20each%20person%20to%20train), haciendo el sistema más robusto a esas variaciones naturales.

## Técnicas de anti‑spoofing en sistemas de biometría de voz (software)

Dado que la voz puede ser grabada o incluso *sintetizada* por imitadores digitales, la biometría vocal enfrenta amenazas de **spoofing** o suplantación. Un atacante podría, por ejemplo, reproducir una grabación de la voz de la víctima (ataque de *replay*) o emplear herramientas de **clonación de voz** (text-to-speech avanzado o conversión de voz) para generar un audio que suene como la persona objetivo. De hecho, con los avances en **deepfakes de voz**, replicar voces humanas de forma realista se ha vuelto factible desde 2017, convirtiendo la **suplantación por voz** en un riesgo creciente[[18]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Perplexingly%2C%20the%20emergence%20of%20deepfakes,past%20a%20voice%20authentication%20check)[[19]](https://www.iproov.com/es/blog/disadvantages-vulnerabilities-voice-biometrics#:~:text=La%20tecnolog%C3%ADa%20biom%C3%A9trica%20de%20voz,expectativas%20de%20rendimiento%20y%20accesibilidad). Por ello, los sistemas biométricos de voz deben incorporar contramedidas **anti-spoofing** que distingan una voz auténtica en vivo de una falsificación. A continuación describimos los principales **mecanismos de software** para anti-spoofing en reconocimiento de voz:

* **Detección de artefactos acústicos en voces sintéticas:** Las voces generadas por algoritmos (por muy reales que parezcan) suelen dejar **huellas sutiles en el espectro** que un sistema puede detectar aunque el oído humano no las note. Por ejemplo, ciertas técnicas de conversión de voz o *text-to-speech* introducen anomalías espectrales en las altas frecuencias o patrones poco naturales en el dominio temporal. Sistemas comerciales como *ID R&D IDLive Voice* utilizan algoritmos capaces de detectar **artefactos espectrales inaudibles** que delatan una voz sintetizada o convertida[[20]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=IDLive%20Voice%20developed%20by%20ID,among%20all%20else%2C%20successfully%20shield). Investigaciones similares han propuesto extraer características espectrales de corto plazo con escalas de frecuencia invertidas o filtrado especial, para resaltar discrepancias entre voces en vivo y voces generadas[[20]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=IDLive%20Voice%20developed%20by%20ID,among%20all%20else%2C%20successfully%20shield). En esencia, se busca identificar inconsistencias en la señal (por ejemplo, ruido de fondo demasiado “perfecto”, falta de micro-fluctuaciones en el timbre, transiciones anómalas entre fonemas, etc.) que indicarían que la muestra no proviene de un humano hablando en tiempo real.
* **Análisis de variabilidad dentro de la misma sesión (detección activa):** Otra estrategia es diseñar el proceso de autenticación de forma que sea difícil de falsificar con una grabación preexistente. Por ejemplo, la solución *Nuance VocalPassword* implementa un desafío aleatorio: el sistema pide al usuario que **repita una frase o parte de ella de forma aleatoria**, capturando dos muestras en la misma sesión[[21]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=VocalPassword%20by%20Nuance%20is%20a,session%20voice%20variation%20principle). Luego compara las dos grabaciones de voz (que deberían ser congruentes si provienen de la misma persona en vivo) y calcula un puntaje de *liveness*. Una grabación estática reproducida por un impostor no podría adaptarse a la petición aleatoria en tiempo real. Esta técnica de *challenge-response* verifica la *viveza* de la voz al introducir espontaneidad. Del mismo modo, algunos sistemas telefónicos pueden solicitar al usuario que diga cierta palabra aleatoria o que cuente números aleatoriamente, dificultando el uso de clips grabados. Estas comprobaciones de coherencia intra-sesión ayudan a **frenar ataques de replay**, aunque por sí solas podrían no detener a una síntesis de voz en tiempo real muy sofisticada (en esos casos se combina con análisis acústico como el anterior).
* **Características de potencia y frecuencia para detección de reproducciones:** Una voz reproducida por un altavoz (ataque de replay) presenta diferencias físicas respecto a una voz humana en directo. Investigadores surcoreanos desarrollaron el sistema **Void**, que analiza la **distribución de la potencia de la señal a lo largo del espectro audible** para detectar patrones anómalos en una posible grabación[[22]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Furthermore%2C%20Void%20computes%20the%20following,FVLPC). En términos prácticos, Void calcula varias características: energía en bandas bajas, grado de linealidad de la potencia, energía en frecuencias altas, coeficientes cepstrales lineales (LPCC), etc., y luego busca **inconsistencias típicas de una reproducción electrónica**[[22]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Furthermore%2C%20Void%20computes%20the%20following,FVLPC). Por ejemplo, un altavoz puede introducir un roll-off en altas frecuencias o una firma de respuesta en frecuencia diferente a la voz humana natural. Combinando estas pistas, Void logró detectar con alta exactitud grabaciones fraudulentas (obteniendo tasas de error iguales –EER– tan bajas como 0.3% en pruebas controladas)[[23]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Void%20was%20tested%20on%20two,learning). Además, al integrar modelos de mezcla gaussiana (GMM) con MFCC, mejoró la detección de voces falsas con buena eficiencia computacional[[23]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Void%20was%20tested%20on%20two,learning). En suma, el análisis de señales regrabadas se apoya en el hecho de que ningún sistema de altavoz/micrófono reproduce la voz con total fidelidad: siempre hay pequeñas pérdidas o distorsiones que un algoritmo de *anti-spoofing* bien entrenado puede aprender a reconocer.
* **Modelos de aprendizaje automático para clasificación real vs. falso:** En la última década, han cobrado fuerza los detectores basados en *machine learning* (redes neuronales, SVM, etc.) entrenados explícitamente para distinguir voces genuinas de voces falsificadas. Estos sistemas suelen alimentarse con **características especializadas anti-spoof** (por ejemplo: coeficientes cepstrales en escala constante - CQCC, características de fase, modulaciones de ruido, etc.) derivadas del audio de entrada, y la red aprende patrones característicos de cada clase (real o impostor). Por ejemplo, en los desafíos internacionales *ASVspoof* se han propuesto múltiples features y arquitecturas de deep learning para detectar *spoofing*. Un modelo típico puede ser una red neuronal convolucional profunda o una LSTM que analice el espectrograma en busca de irregularidades. La ventaja de estos enfoques es que combinan múltiples pistas (timbre, dinámica, microdetalle) de forma no lineal y a menudo alcanzan altos porcentajes de detección. No obstante, requieren grandes conjuntos de entrenamiento con ejemplos de ataques conocidos, y pueden ser burlados por ataques de tipo nuevo no vistos durante el entrenamiento. Aun así, constituyen una capa de seguridad esencial: por ejemplo, *ID R&D* y otros líderes del sector reportan tasas de detección excelentes en evaluaciones públicas, gracias al uso de redes neuronales que detectan *artefactos* dejados por los sintetizadores de voz más avanzados[[20]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=IDLive%20Voice%20developed%20by%20ID,among%20all%20else%2C%20successfully%20shield)[[24]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Image%3A%20ID%20R%26D%E2%80%99s%20performance%20during,off).
* **Otras técnicas de verificación activa (señales auxiliares):** Si bien la pregunta se centra en mecanismos de software (acústicos), vale la pena mencionar brevemente que existen métodos que aprovechan el hardware estándar de los dispositivos para reforzar la detección de vida. Por ejemplo, esquemas como **VoiceGesture** utilizan el altavoz del teléfono para emitir un tono ultrasónico (~20 kHz) durante la autenticación; este tono inaudible rebota en los movimientos articulatorios del hablante y el micrófono capta los **cambios Doppler** producidos[[25]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=VoiceGesture%20is%20based%20on%20the,profile%2C%20liveness%20will%20be%20confirmed). Como el patrón de variación Doppler es distinto si la voz proviene de un ser humano moviendo la boca versus de un altavoz fijo reproduciendo audio, el sistema puede confirmar si hay una persona presente. Otro enfoque investigado es medir ligeras variaciones de presión en el canal auditivo del usuario cuando habla (lo cual requeriría sensores específicos) para comprobar liveness. Incluso se han explorado técnicas que analizan el tiempo de llegada de la voz a múltiples micrófonos para diferenciar una fuente humana cercana de un altavoz distante. Aunque estos métodos van más allá del simple análisis de la onda de audio, todos comparten el objetivo de **imposibilitar que una simple reproducción engañe al sistema**. En aplicaciones de alto riesgo, una combinación de múltiples capas – desde análisis acústico puro hasta desafíos activos y sensores adicionales – brinda la mayor seguridad contra spoofing.

**En resumen**, un sistema de reconocimiento de voz robusto no solo extrae una amplia gama de **características biométricas de la voz** (timbre, tono, formantes, etc.) para identificar correctamente al hablante bajo diversas condiciones (incluso si tiene una leve gripa), sino que también implementa **contramedidas de software anti-spoofing** para asegurarse de que la voz proviene de la persona real en el momento de la autenticación. La combinación de estas estrategias permite que la biometría vocal sea fiable y segura, dificultando suplantaciones incluso frente a las crecientes amenazas de voces sintéticas ultra-realistas[[18]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Perplexingly%2C%20the%20emergence%20of%20deepfakes,past%20a%20voice%20authentication%20check). Los continuos avances en procesamiento de señales e inteligencia artificial siguen mejorando tanto la **precisión** de la identificación por voz como la **detección temprana de fraudes**, consolidando la voz como un medio práctico de autenticación en entornos cada vez más vigilados.

**Fuentes:** Las características vocales aquí descritas y su relevancia se han recopilado de literatura especializada y proyectos de biometría de voz[[26]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identidad,Tono%20%28%20pitch)[[27]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,Miden%20caracter%C3%ADsticas%20espec%C3%ADficas%20de)[[28]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,partir%20de%20la%20descomposici%C3%B3n%20arm%C3%B3nica), mientras que las técnicas anti-spoofing se han basado en reportes actuales de la industria y estudios académicos sobre detección de falsificación de audio[[20]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=IDLive%20Voice%20developed%20by%20ID,among%20all%20else%2C%20successfully%20shield)[[21]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=VocalPassword%20by%20Nuance%20is%20a,session%20voice%20variation%20principle)[[22]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Furthermore%2C%20Void%20computes%20the%20following,FVLPC), entre otros. Estas referencias respaldan las definiciones y métodos presentados, proporcionando un panorama actualizado (2024-2025) de las mejores prácticas en **biometría vocal** y seguridad anti-spoofing.

[[1]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=concentrates%20on%20choosing%20the%20proper,can%20be%20identified%20or%20verified) [[2]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=Automatic%20speaker%20recognition%20is%20an,physical%20state%2C%20and%20emotional%20state) [[3]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=%60%60Cold,from%20each%20person%20to%20train) [[4]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=the%20normal%2C%20healthy%20signals,the%20acoustic%20signal%20and%20the) [[17]](https://acoustics.org/pressroom/httpdocs/131st/lay10a.html#:~:text=The%20goal%20of%20this%20research,physical%20state%2C%20and%20emotional%20state) Acoustical Society of America - ``Cold Speech'' for Automatic Speaker Recognition

<https://acoustics.org/pressroom/httpdocs/131st/lay10a.html>

[[5]](https://www.iproov.com/es/blog/disadvantages-vulnerabilities-voice-biometrics#:~:text=dificultando%20o%20imposibilitando%20a%20algunas,hacer%20coincidir%20la%20huella%20vocal) [[19]](https://www.iproov.com/es/blog/disadvantages-vulnerabilities-voice-biometrics#:~:text=La%20tecnolog%C3%ADa%20biom%C3%A9trica%20de%20voz,expectativas%20de%20rendimiento%20y%20accesibilidad) Desventajas y vulnerabilidades de la biometría de voz | Buscando alternativas | iProov

<https://www.iproov.com/es/blog/disadvantages-vulnerabilities-voice-biometrics>

[[6]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,resonantes%20que%20caracterizan%20el%20timbre) [[7]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,Root%20Mean) [[8]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=%2F%20Segundo%20semestre%202024%2056,permite%20capturar%20la%20esencia%20de) [[10]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,dispersi%C3%B3n%20de%20las%20frecuencias%20en) [[11]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identi,Tono%20%28%20pitch) [[12]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Formantes%20%28F1%2C%20F2%2C%20F3%2C%20F4%29,frecuencia%20a%20lo%20largo%20del) [[13]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Square%20%29,variabilidad%20en%20el%20periodo%20y) [[16]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=Square%20%29,Miden%20caracter%C3%ADsticas%20espec%C3%ADficas%20de) [[26]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=identidad,Tono%20%28%20pitch) [[27]](https://www.segurilatam.com/revistas/ciberilatam/002/56/#:~:text=complejidad%20y%20variabilidad%20del%20es,Miden%20caracter%C3%ADsticas%20espec%C3%ADficas%20de) Ciberilatam 002

<https://www.segurilatam.com/revistas/ciberilatam/002/56/>

[[9]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=En%202019%2C%20se%20desarroll%C3%B3%20una,1) [[14]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,del%20tono%20durante%20la%20conversaci%C3%B3n) [[15]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,partir%20de%20la%20descomposici%C3%B3n%20arm%C3%B3nica) [[28]](https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/#:~:text=,partir%20de%20la%20descomposici%C3%B3n%20arm%C3%B3nica) Identificación biométrica de personas a través de la voz – PITCH

<https://pitch.xeridia.com/identificacion-biometrica-de-personas-a-traves-de-la-voz/>

[[18]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Perplexingly%2C%20the%20emergence%20of%20deepfakes,past%20a%20voice%20authentication%20check) [[20]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=IDLive%20Voice%20developed%20by%20ID,among%20all%20else%2C%20successfully%20shield) [[21]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=VocalPassword%20by%20Nuance%20is%20a,session%20voice%20variation%20principle) [[22]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Furthermore%2C%20Void%20computes%20the%20following,FVLPC) [[23]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Void%20was%20tested%20on%20two,learning) [[24]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=Image%3A%20ID%20R%26D%E2%80%99s%20performance%20during,off) [[25]](https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/#:~:text=VoiceGesture%20is%20based%20on%20the,profile%2C%20liveness%20will%20be%20confirmed) Voice liveness detection: methods overview — Antispoofing Wiki

<https://antispoofing.org/voice-liveness-detection-systems-challenges-and-solutions/>