APROXIMACIÓN AL ANÁLISIS COMPARATIVO DE VOZ ARTIFICIAL MEDIANTE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTORES

APPROACH TO COMPARATIVE ANALYSIS OF ARTIFICIAL VOICE THROUGH AN AUTOMATIC SPEAKER RECOGNITION SYSTEM

Ana MOLINERO ESCAPA

Dña. María Ángeles MARTÍN RUBIO
Sección de Acústica Forense
Comisaría General de Policía Científica

Dña. María Concepción ALONSO RODRÍGUEZ
Departamento de Física y Matemáticas
Universidad de Alcalá de Henares

MÁSTER UNIVERSITARIO EN CIENCIAS POLICIALES

(Trabajo fin de Máster – 12 ECTS)

2020-2021

INSTITUTO UNIVERSITARIO DE INVESTIGACIÓN EN CIENCIAS POLICIALES

Agradecimientos

El Covid-19 ha marcado nuestra vida, resaltando tanto nuestra fragilidad como nuestra fortaleza, capacidad de adaptación y superación. En este escenario tan desconocido y complejo he desarrollado este Trabajo Fin de Máster (TFM) con la ayuda y orientación de muchas personas a las que quiero dar las gracias de corazón. Soy muy consciente de su esfuerzo y dedicación:

A D. Carlos Delgado Romero, Jefe de la Sección de Acústica Forense de la Comisaría General de Policía Científica (CGPC), por alentarme a realizar el Máster y por sus clases magistrales.

A Dña. María Ángeles Martín Rubio, mi tutora y maestra durante las prácticas, por su ayuda y enseñanza. Sin ella no hubiese sido posible culminar este trabajo. Gracias infinitas por su paciencia, disponibilidad y explicaciones, su empatía y bondad.

A Dña. María Concepción Alonso Rodríguez, cotutora y Directora del Máster Universitario en Ciencias Policiales, por su disposición y ánimo durante la elaboración de este TFM.

A D. Javier Méndez Nevado por su tiempo y sencillez en la explicación de difíciles conceptos matemáticos.

A Dña. María Rosario Lafuente Avilés y resto del equipo del laboratorio de Acústica por compartir conmigo sus conocimientos y experiencia.

A D. Luis Hernández-Hurtado García, responsable institucional del Máster Universitario en Ciencias Policiales, Comisario Jefe de la Unidad Central de Investigación Científica y Técnica de la CGPC donde presto servicio, por impulsar la formación del personal policial.

Y, por último, quiero agradecer a la Secretaría de Estado de Seguridad, Dirección General de la Policía y Comisaría General de Policía Científica por facilitarme el acceso a estos estudios.

Gracias a todos por hacerlo posible.

Índice

Resumen

El ritmo de vida acelerado al que nos vemos sometidos en los países desarrollados ha llevado a las empresas a dar respuesta a una necesidad de uso y control de *software* y *hardware* mucho más efectiva. Es así como se incorpora la utilización de la voz para facilitar cualquier comunicación cotidiana, desde consultar el tiempo que va a hacer durante el día hasta la interacción con sistemas sofisticados que utilizan la inteligencia artificial para realizar tareas más complejas como solicitar la lectura de un texto.

En este último punto se centra el presente trabajo, que aborda el estudio comparativo de muestras de voz sintética, obtenidas a partir de tres aplicaciones gratuitas (SPIK-AI, NUANCE VOCALIZER y Play HT), utilizando BATVOX 4.1, el Sistema de Reconocimiento Automático de Locutores usado por la mayoría de laboratorios de Acústica Forense de todo el mundo. Nuestro objetivo es evaluar su capacidad de discriminación frente a este tipo de locuciones y determinar si los resultados alcanzados tienen una validez suficiente para considerar su utilización.

El experimento realizado revela, por un lado, que la mayoría de muestras de voz artificial no cumplen con los requisitos requeridos por el sistema, bien debido a su formato de audio o bien por los desajustes con las poblaciones de referencia disponibles. Por otro lado, para las muestras útiles, aunque los resultados no son incoherentes, se observa que la capacidad de discriminación no es del todo adecuada por lo que no es recomendable su uso con este tipo de habla.

Palabras clave: Acústica Forense, análisis comparativo de voces, voz sintética, redes neuronales artificiales, Sistema de Reconocimiento Automático de Locutores, modelo de locutor, audio test, audio de impostor.

Abstract

The accelerated life to which we are subjected in the more developed countries has led companies to respond to a need for the use and control of software and hardware that is much more effective. This is how the use of the voice is incorporated to facilitate any daily communication, from consulting the weather that is going to be done during the day to interacting with sophisticated systems that use artificial intelligence to perform more complex tasks such as requesting a reading.

This last point is the focus of this work, which addresses the comparative study of synthetic voice samples, obtained from three free applications (SPIK-AI, NUANCE VOCALIZER and Play HT), using BATVOX 4.1, the Automatic Speaker Recognition System used by the majority of Forensic Acoustics laboratories around the world. Our objective is to evaluate their ability to discriminate against this type of utterance and determine if the results achieved have sufficient validity to consider their use.

The experiment carried out reveals, on the one hand, that the majority of artificial voice samples do not meet the requirements required by the system, either due to their audio format or due to mismatches with the available reference populations. On the other hand, for useful samples, although the results are not inconsistent, it is observed that the discrimination capacity is not entirely adequate, so its use with this type of speech is not recommended.

Keywords: Forensic Acoustics, comparative voice analysis, synthetic voice, artificial neural networks, Automatic Speaker Recognition System, speaker model, audio test, impostor audio.

1. Introducción.

En la última década la tecnología informática ha experimentado un gran avance, siendo en los teléfonos inteligentes, los *smartphones*, donde las empresas han realizado mayores inversiones dirigidas a aumentar la memoria, a mejorar la definición de la pantalla o la durabilidad de la batería. En paralelo a este crecimiento, los desarrolladores han creado aplicaciones informáticas que permiten llevar a cabo una gran diversidad de tareas. Con un solo clic y de manera gratuita, el usuario puede disponer de prácticamente cualquier herramienta.

La delincuencia siempre ha ido en consonancia con la tecnología y la voz, clave en los procesos de comunicación, no se sustrae a esta realidad. Hace tiempo que no resulta extraño que se utilice la voz con el fin, por ejemplo, de obtener un beneficio económico. En nuestro país, aún resuena el caso de Anabel Segura, aquella joven que fue secuestrada y asesinada por sus captores. Sus asesinos hicieron creer a la familia de la víctima que esta permanecía con vida imitando su voz para poder cobrar el rescate solicitado. En la actualidad, el malhechor podría acudir al mercado virtual y servirse de lo que ofrecen las nuevas tecnologías del habla. Su evolución en las últimas décadas nos ha llevado desde las primeras voces sintéticas con timbre metalizado usadas en las máquinas expendedoras hasta los asistentes por voz que ofertan las grandes multinacionales de la comunicación (Amazon, Apple, Google…); algunas transacciones en banca se pueden realizar gracias al reconocimiento de voz y, en los límites con la ciencia ficción, la clonación de voz permite generar grandes cantidades de habla de una persona a partir de tan solo unos segundos de su voz natural (a todos nos sorprendió oír la voz de Lola Flores en el anuncio de CruzCampo [1]).

Dejando aparte los valores que cada uno posea y el uso que cada cual haga de estas aplicaciones informáticas, la investigación forense debe centrar sus esfuerzos en combatir una realidad hoy en día ineludible: la utilización de voz sintética con fines delincuenciales. El abanico de este uso ilícito puede ser amplio: fraudes financieros, extorsiones, secuestros, amenazas, terrorismo, usurpación de identidad, delitos contra la intimidad, etc.

El presente estudio, “Aproximación al análisis comparativo de voz artificial mediante un Sistema de Reconocimiento Automático de Locutores”, como culminación del Máster Universitario en Ciencias Policiales en colaboración con la Comisaría General de Policía Científica, abordará varios ensayos en laboratorio con voces sintéticas generadas por aplicaciones informáticas de libre acceso que utilizan inteligencia artificial basada en redes neuronales, con el fin de determinar si las herramientas forenses de Reconocimiento Automático de Locutores disponibles son capaces de discriminar entre diferentes voces de este tipo.

Para ello, se han generado grabaciones de voz sintética con tres aplicaciones de conversión texto a voz, SPIK-AI, NUANCE VOCALIZER y Play HT, con duraciones y características de registro adecuadas a los requerimientos del Sistema de Reconocimiento Automático utilizado. Se han realizado las comparaciones entre las distintas voces de varón disponibles en una misma aplicación y por separado para cada una de las tres aplicaciones mencionadas. Los resultados obtenidos permiten determinar que la forma en la que se generan las grabaciones de voz artificial afecta de manera importante al funcionamiento del Sistema de Reconocimiento Automático de Locutores en su versión actual, por lo que su uso en exclusiva no es aconsejable con este tipo de grabaciones.

2. Fundamentos teóricos del sonido. La voz y el habla.

2.1. Fundamentos teóricos del sonido. Formas de representación gráfica.

La acústica es la ciencia del sonido, y este, expresado de una forma muy simple, es aire en movimiento que nuestro oído, como si fuese un micrófono, capta y transforma en información que se descodifica en nuestro cerebro.

Desde el punto de vista físico, un sonido es una perturbación producida en las partículas de un medio elástico por un cambio de presión que se transmite a través del mismo en forma de onda mecánica. Si además tenemos en cuenta la interacción de ese sonido con los mecanismos auditivos del ser humano, este deberá encuadrarse en valores de intensidad y frecuencia delimitados para ser audible; concretamente

podemos percibir aquellos sonidos con frecuencias entre los 20 y los 20000 Hz siempre y cuando su intensidad esté por encima de un umbral determinado. El habla humana, como un tipo especial de sonido, transcurre entre los 100 y los 7000 Hz [2,3].

En la gráfica siguiente se representa la intensidad frente a la frecuencia con los umbrales mencionados. Dentro del rango establecido, a altas y bajas frecuencias, necesitaremos que el sonido tenga mayor intensidad para poder escucharlo. Lo contrario ocurre para una frecuencia de 3,5 Hz aproximadamente (la equivalente al llanto de un bebé): se requiere muy poca intensidad para poder percibirlo.

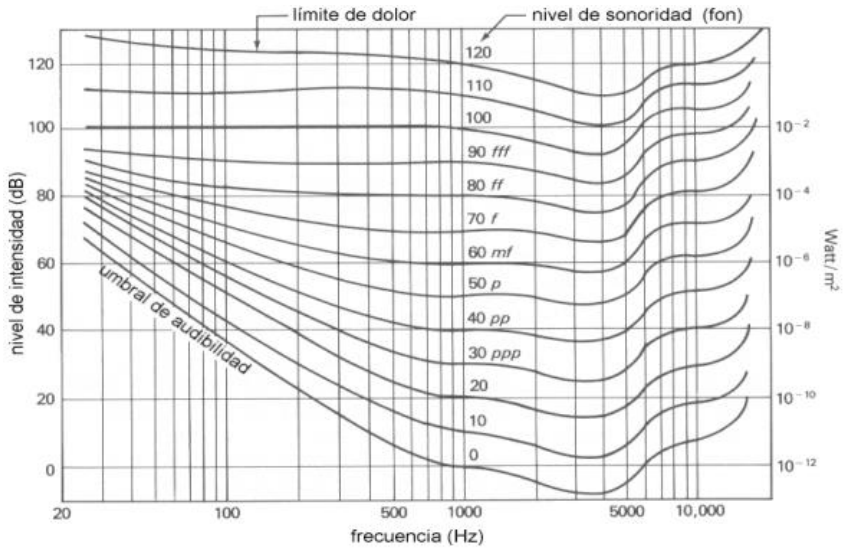

Figura 1. Gráfica que representa el umbral de audición del ser humano.

Básicamente, el sonido se dimensiona en torno a tres ejes, las tres magnitudes físicas que lo caracterizan:

- La **frecuencia** (f): número de vibraciones por segundo de las partículas del medio. La unidad de medida de la frecuencia es el hercio (Hz) o número de ciclos por segundo, un hercio equivale a un ciclo por segundo.
- El **tiempo** (t). Es decir, cuánto dura el sonido analizado, medido en segundos (s).

- La **intensidad** (I) o presión acústica, se mide en decibelios (dB) y nos informa acerca de la amplitud del desplazamiento de las partículas del medio respecto de su posición de equilibrio.

Teniendo en cuenta la componente de frecuencia podemos clasificar los sonidos en simples (aquellos que presentan una única frecuencia) y complejos (en los que aparecen varias frecuencias). Los sonidos simples o tonos puros se representan matemáticamente mediante una función sinusoidal. En la naturaleza no se dan los tonos puros sino los sonidos formados por la adición de varios tonos simples. El análisis de Fourier permite descomponer una onda compleja con carácter periódico en varias ondas simples denominadas armónicos. Según el estudio de Fourier, las frecuencias de estos armónicos son múltiplos enteros del primero de ellos o frecuencia fundamental que, además, coincide con la de la onda compleja. Este proceso se muestra en la figura 2.

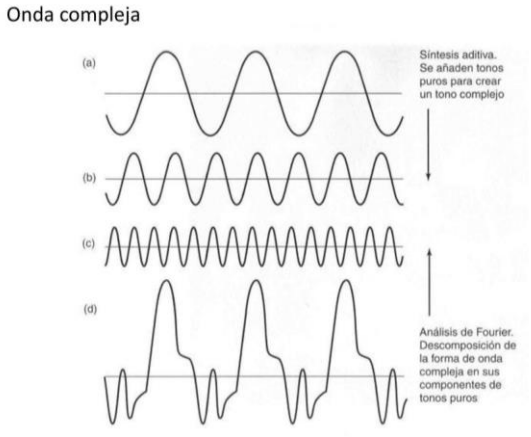

Figura 2. Representación gráfica de la síntesis y descomposición de una onda compleja.

Este concepto es interesante para nuestro trabajo, ya que el proceso inverso, la integración de diferentes ondas de sonido para obtener una onda compleja- lo que se conoce como síntesis de Fourier-, es el procedimiento utilizado por algunas aplicaciones digitales específicas para generar voz sintética.

En el análisis de los sonidos del habla es fundamental el efecto que se produce cuando una onda compleja se transmite a través de un objeto con forma tubular: el

fenómeno de la resonancia, por el cual se produce una amplificación de aquellas frecuencias que coincidan con las de resonancia del tubo, disminuyendo la intensidad del resto de frecuencias. Esto nos proporciona un nuevo eje que podemos añadir a la frecuencia, el tiempo y la amplitud: la estructura acústica de resonancia, que permite distinguir los sonidos en función de la forma que les confiere las características de la caja de resonancia en la que se han producido. Es el caso de los sonidos del habla generados, como veremos más adelante, utilizando las cavidades que conforman el tracto vocal.

Una parte importante del estudio de la voz se basa en el uso de las distintas formas de representación gráfica del sonido. En función de las magnitudes utilizadas se obtienen distintos tipos de representación. Así, podemos obtener una forma de onda u oscilograma, un espectro o un espectrograma o sonograma.

El **oscilograma** representa los valores de intensidad en función del tiempo. Un registro sonoro quedaría representado en forma de oscilograma como se muestra en la figura 3. En la ventana superior aparece un oscilograma de unos 19 segundos de duración y en la ventana inferior se ha realizado un *zoom* sobre el anterior, mostrando un tramo de periodicidad que podría corresponderse con un sonido vocálico, por ejemplo.

Figura 3. Oscilogramas o formas de onda.

La misma onda sonora que hemos representado en función del tiempo, podemos representarla en el dominio de la frecuencia mediante la Transformada de Fourier (TF), operador matemático desarrollado a partir del análisis de descomposición expuesto anteriormente. Obtendríamos gráficamente lo que se denomina **espectro**.

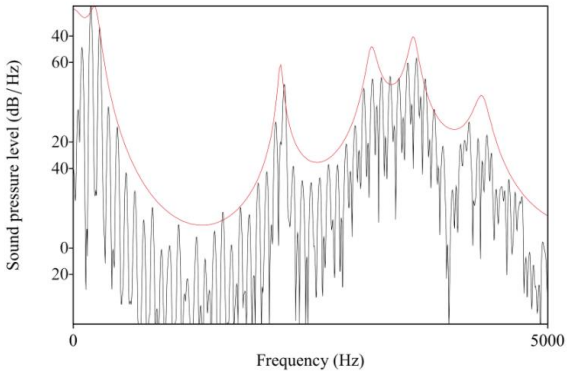

Figura 4. Espectro correspondiente a la realización de un sonido /i/.

Por último, el **espectrograma o sonograma** (figura 5) ofrece una imagen tridimensional del sonido: representa la frecuencia en función del tiempo mostrando los niveles de intensidad de las distintas frecuencias mediante una escala de grises o de colores. Es la representación gráfica que mejor permite visualizar las características y peculiaridades de los diferentes sonidos del habla, como veremos más adelante.

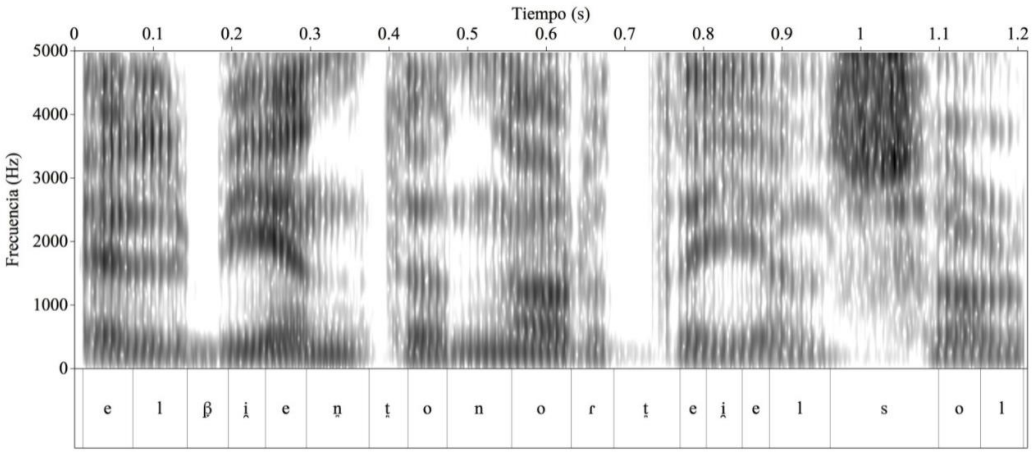

Figura 5. Sonograma del enunciado "El viento norte y el sol".

2.2. Fase productora del habla.

En la producción del habla intervienen varios grupos de órganos: de manera indirecta, los órganos respiratorios (bronquios, pulmones y tráquea) que generan con el aire espirado la presión subglótica suficiente para producir la vibración de las cuerdas vocales; y los órganos implicados directamente en la producción de la voz, la cavidad laríngea, con la laringe y las cuerdas vocales y las cavidades resonadoras (faringe, cavidad bucal y cavidad nasal) [4,5].

La figura 6 muestra una sección del tracto vocal con los diferentes elementos que intervienen en la realización de las emisiones habladas.

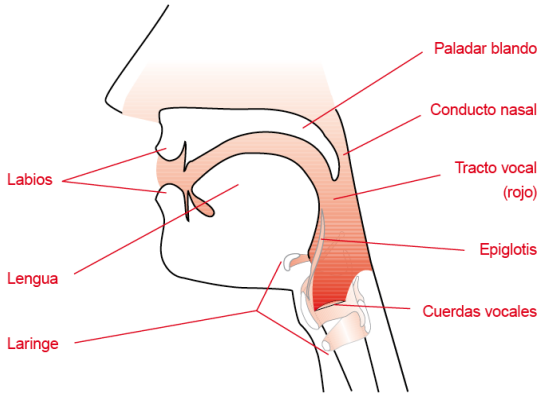

Figura 6. Sección del tracto vocal.

El aire expulsado proveniente de los pulmones se encuentra en la cavidad laríngea con el principal responsable de la producción del habla, las cuerdas vocales, dos bandas de tejido muscular flexible que tienen la capacidad de vibrar generando así una onda compleja de carácter periódico que dará lugar a sonidos con ese carácter de periodicidad (sonidos sonoros). Dicha onda estará formada por un primer armónico, denominado frecuencia fundamental (F0) y los armónicos con frecuencias múltiplos de la misma. La onda generada atraviesa las cavidades resonadoras que, en función de la colocación de los órganos que intervienen en la articulación (paladar, lengua, dientes, labios) adquiere una forma u otra dando lugar a sonidos distintos (figura 7).

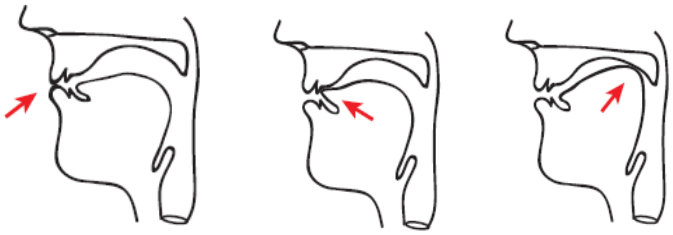

Figura 7. Distintas posiciones de los articuladores.

La acción de las cavidades resonadoras sobre la onda original generada a nivel glotal se traduce en una estructura acústica de resonancia que dependerá de la fisonomía particular de cada individuo, así como de los hábitos articulatorios adquiridos. Aquellas frecuencias que el resonador ha amplificado dan lugar a lo que se denomina formantes.

Por tanto, en el estudio de una emisión hablada podemos distinguir dos informaciones: la información a nivel de vibración de las cuerdas vocales, F0, y la que proviene de la acción del tracto vocal, la estructura de resonancia. Esta última, por tanto, es la más útil a la hora de discriminar entre dos voces distintas ya que contiene las características que la anatomía particular de cada individuo confiere a sus emisiones habladas.

En la figura siguiente se muestran las gráficas correspondientes a la emisión “era tarde” en su forma de onda (ventana A), en representación espectrográfica o sonograma (ventana B) y el espectro correspondiente al sonido vocálico /e/ (ventana C).

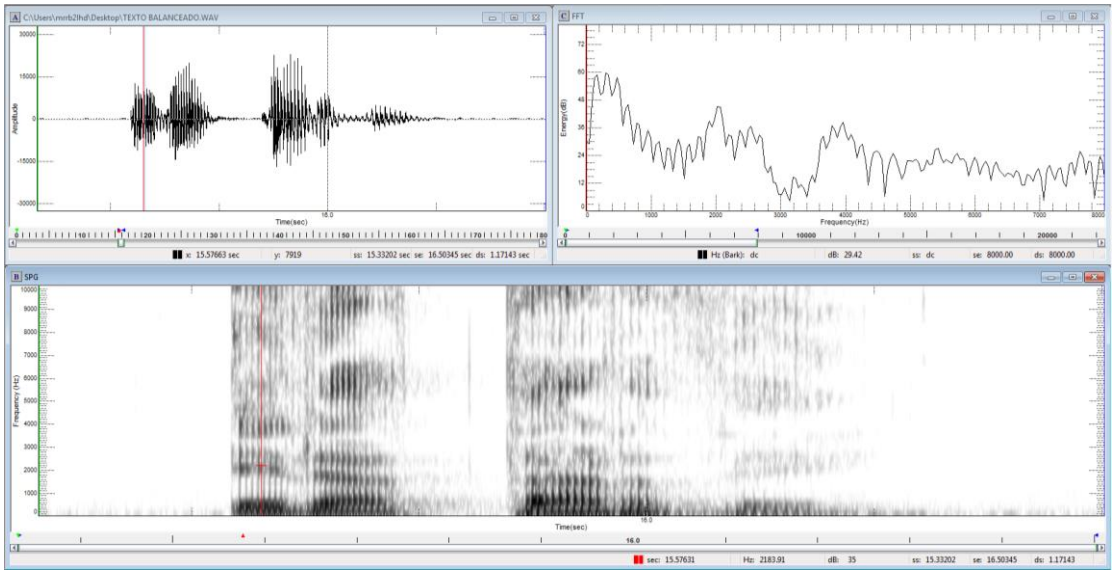

Figura 8. Distintas representaciones gráficas de una emisión hablada.

Las zonas más oscuras que se observan en el sonograma son los **formantes**: picos de mayor intensidad que se corresponden con las resonancias y que variarán para cada locutor en función de la anatomía de su tracto vocal y de su estructura craneal. Igualmente, un mismo locutor emitirá distintos sonidos, con estructura formántica diferente según la posición de los articuladores.

Además de los sonidos sonoros en los que se produce una vibración de las cuerdas vocales podemos emitir sonidos en los que esto no sucede, sonidos sordos que, desde un punto de vista puramente acústico, se conforman como ruido. En la figura 9 se muestra la estructura acústica de la frase “casi de noche”; en el sonograma representado en la ventana B los cursores azules delimitan la realización del sonido sordo /s/, que como se aprecia no presenta la estructura de formantes propia de los sonidos sonoros.

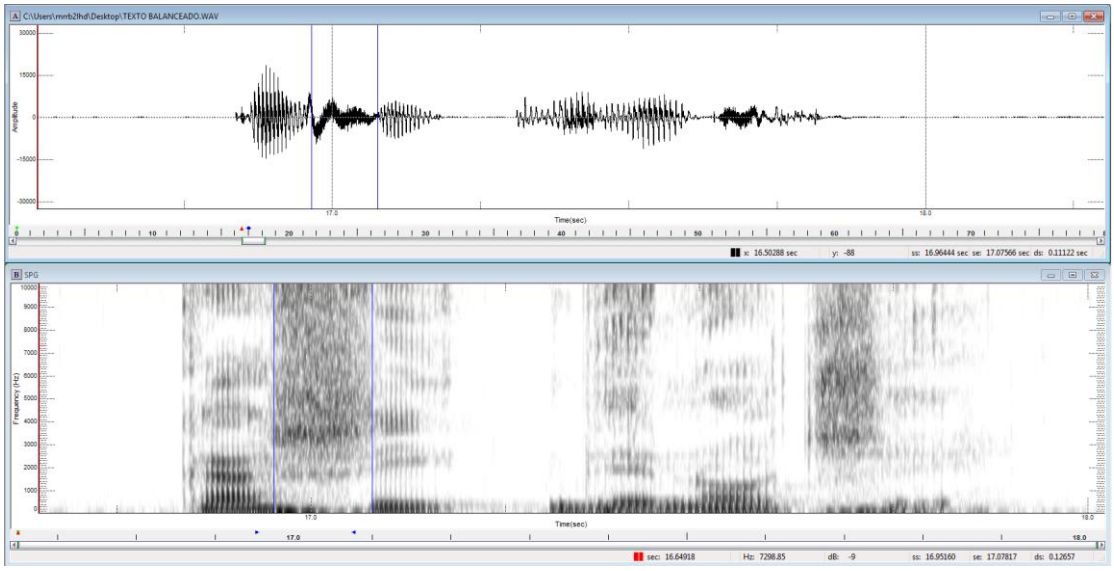

Figura 9. Oscilograma y sonograma de la frase “casi de noche” en los que aparece delimitado el sonido /s/.

2.3. Conceptos básicos de psicoacústica.

Los sonidos del habla que llegan a nuestro oido en forma de estímulos auditivos producen, desde un punto de vista psicológico, un determinado grado de sensación. Así, las propiedades puramente físicas de un sonido se transforman, mediante el proceso de la percepción, en apreciaciones de tipo subjetivo que dependerán de las caracterísitcas y capacidades del sujeto perceptor.

Desde una perspectiva psicoacústica podemos establecer una correspondencia entre las magnitudes físicas que caracterizan el sonido y la forma en la que las percibimos [2]:

- La frecuencia se correlaciona con lo que denominamos **tono**: lo que percibimos como un tono bajo, una voz grave, se corresponde con un valor de F0 también bajo. Y, por el contrario, un sonido percibido como agudo presentará una frecuencia alta. La relación entre la frecuencia real de un sonido y la frecuencia percibida (tono) solo es lineal para valores hasta los 1000 Hz. A partir de este valor la relación responde a una escala logarítmica como se muestra en la gráfica de la figura 10. Esta escala recibe el nombre de escala MEL.

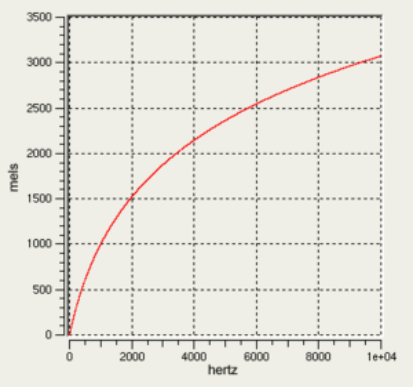

Figura 10. Escala MEL de frecuencia.

- La intensidad tiene su correlato perceptivo en la **sonoridad**, directamente relacionada con la amplitud de la onda sonora.

- La **duración** de un sonido es lo que subjetivamente nos parece que abarca ese sonido en el tiempo.

- El correlato psicoacústico de la estructura de resonancia es lo que se denomina **timbre** o **cualidad vocal** y es la propiedad que, fundamentalmente, nos permite diferenciar unas voces de otras.

2.4. Factores de variabilidad.

La principal característica de la voz, que va a condicionar de manera relevante la forma en la que debe abordarse su estudio, es su variabilidad [6]. Se trata de una referencia biométrica dinámica de la conducta humana, como la forma de andar o escribir. Este carácter variable implica que en los análisis comparativos de voces, para determinar si un acto de habla puede haber sido realizado por una persona determinada, es necesario tener en cuenta no sólo la variabilidad interpersonal sino también la intrapersonal, pues no existen dos actos de habla absolutamente idénticos.

Existen numerosos factores que pueden incidir en esta variabilidad intrapersonal:

- Contemporaneidad de las muestras: está comprobado que, en un lapso de tiempo de diez años, la frecuencia fundamental de una voz (en relación

directa con la frecuencia de vibración de las cuerdas vocales) varía. En términos generales, se produce un aumento de la misma en el caso de los varones y una disminución en las mujeres, debido a las transformaciones de la musculatura de las cuerdas vocales por efecto de los cambios hormonales.

- Entorno anatómico: incorporación de prótesis dentarias, formación de nódulos, etc.
- Entorno fisiológico: aparición de patologías del habla, disfonías, procesos inflamatorios, etc.
- Entorno psicológico: temblor corporal, situación emocional, etc.
- Variaciones anímicas como la tristeza, la excitación, temor, etc.
- Influencia de agentes químicos: uso de medicamentos, sustancias estupefacientes, alcohol, tabaco...

3. Áreas de trabajo en el laboratorio de Acústica Forense de la Policía Nacional. Metodología de trabajo en la identificación de locutores. Sistema de Reconocimiento Automático de Locutores (SRAL).

3.1. Áreas de trabajo en el laboratorio de Acústica Forense.

En el conjunto de las disciplinas que colaboran con la justicia en el esclarecimiento de un hecho delictivo aportando pruebas o indicios de diferente naturaleza, se encuentra la Acústica Forense, cuyo objeto de estudio es el sonido y, fundamentalmente, un tipo especial de sonido, la voz. En el camino hacia la identificación del autor de un hecho delictivo, cuando aparecen actos del habla que pueden tener relación con el mismo, el órgano judicial competente puede solicitar la realización de diferentes tipos de análisis en el campo de la Acústica Forense.

El laboratorio de Acústica Forense del Cuerpo Nacional de Policía pertenece a la Unidad Central de Criminalística de la Comisaría General de Policía Científica. Es el

único laboratorio de este tipo con el que cuenta la Policía Nacional por lo que su ámbito de actuación es a nivel nacional.

Actualmente está formado por dos departamentos integrados por personal policial, técnico y facultativo con formación multidisciplinar (filólogos, físicos, técnicos de sonido, psicólogos) desarrollando las áreas de trabajo que se enumerarán más adelante.

Los estudios llevados a cabo pueden iniciarse tanto a requerimiento de la autoridad judicial competente como a instancias de otros órganos policiales. Las muestras, en sus correspondientes soportes analógicos o digitales, se registran en la base de datos policiales BINCIPOL, asegurando así la adecuada cadena de custodia durante el análisis y hasta la remisión de las mismas junto con el informe emitido.

- **Análisis comparativo del habla:** Es el análisis más demandado, tanto por los órganos judiciales como por las unidades policiales de investigación. El objeto del mismo es el cotejo de grabaciones de voz con el fin de determinar si los actos de habla registrados en las mismas pueden haber sido realizados o no por la misma persona.

  Las muestras de carácter dubitado, consistentes en grabaciones de voz atribuida, pueden provenir de intervenciones telefónicas, de grabaciones directas con un teléfono móvil u otros sistemas de registro, de grabaciones de video, etc.

  En una primera instancia se valorará la calidad y cantidad de dichas muestras. En el ámbito forense, las grabaciones suelen presentar elementos de degradación (distorsiones, ruidos, solapamiento de voces, etc.) que pueden dificultar o incluso impedir la realización de los estudios requeridos.

  El laboratorio establece un protocolo de calidad que estipula los requisitos mínimos que deben presentar las muestras para su admisibilidad. De acuerdo a este protocolo el experto determinará si las muestras son aptas o no para el análisis, así como los condicionantes del mismo, emitiendo el correspondiente informe.

Si este resulta favorable, se procederá al estudio comparativo de la voz dubitada con la voz indubitada del investigado, tomada generalmente en sede judicial por los mismos funcionarios que llevan a cabo los análisis.

Se trata de una técnica compleja dado que, como se ha explicado en el anterior epígrafe, el habla es una referencia biométrica variable, se puede modificar voluntaria o involuntariamente, puede verse afectada por agentes endógenos (estado anímico o emocional de la persona, envejecimiento, patologías, etc.) o exógenos (acción de agentes químicos o padecimiento de enfermedades).

Por ello en el análisis comparativo de voces se aplica lo que se denomina "método combinado" que, como desarrollaremos en el epígrafe 3.2, aúna diferentes perspectivas de estudio: de percepción auditiva, acústica-espectrográfica, fonético/lingüística y de reconocimiento automático.

- **Procesado de señal**: consiste en la edición y filtrado de la señal acústica con el objetivo de mejorar la inteligibilidad de locuciones o aislar eventos de registro.

- **Pasaporte vocal:** Este estudio se realiza a petición del personal policial operativo que realiza la investigación en aquellos casos en los que solo se dispone de una grabación de voz dubitada perteneciente a un sospechoso del que se desconoce la identidad. Los expertos del laboratorio analizan las emisiones habladas del sospechoso trazando un perfil lingüístico del mismo determinando, siempre de manera orientativa, factores socioculturales, rango de edad, género, procedencia geográfica, posibles patologías, etc.

- **Autentificación:** Se realizan análisis del soporte, tanto analógico como digital, para descartar la existencia de una posible de manipulación en la grabación o si ésta se ha realizado en un lugar o con un dispositivo en concreto.

- **Ruedas de reconocimiento de voz:** Realizadas en sede judicial en aquellos casos en que existen víctimas o testigos del hecho delictivo que han

escuchado la voz del autor. Se confeccionan cadenas habladas utilizando voces de similares características a la del sospechoso con el fin de que la víctima o testigo realice un reconocimiento auditivo.

3.2. Metodología de trabajo en identificación de locutores en el laboratorio de Acústica Forense de la Policía Nacional.

La técnica de identificación de locutores comenzó a desarrollarse a finales del siglo XX con el auge de la tecnología, pero desde finales del siglo XVII existen numerosas referencias bibliográficas donde ya se admitía la prueba de identificación de personas a través de su voz en sede judicial, teniendo en consideración el reconocimiento perceptivo por parte de la víctima.

A partir de los años 50, ingenieros estadounidenses fueron perfeccionando la técnica, consiguiendo codificar el habla en formas gráficas utilizando el sonógrafo. Esta herramienta novedosa permitió representar el sonido hablado en tres dimensiones a partir de los parámetros de frecuencia, amplitud y tiempo.

En los años siguientes el físico Lawrence G. Kersta desarrolló un método de identificación basado en el uso de forma exclusiva del análisis sonográfico, asimilando la comparación de lo que denominó *voiceprint* (huella de voz) al cotejo de las impresiones dactilares. Esto terminó en fracaso; su gran error fue desestimar la distinta naturaleza de ambos objetos de estudio, el carácter variable de las emisiones habladas frente a la inmutabilidad de la huella dactilar. Sin embargo, sentó de alguna manera las bases para la incorporación de la perspectiva auditiva a los métodos espectrográficos.

Paralelamente al desarrollo de esta técnica, que se fue extendiendo por el resto de países, incorporándose a los recién creados laboratorios de acústica de Europa y Asia, otros investigadores centraron su metodología en el uso de la fonética, rechazando el análisis espectrográfico.

El enfrentamiento entre ingenieros, defensores de los métodos espectrográficos, y fonetistas continuó durante las siguientes décadas.

En los 70, el físico Oscar Tosi, asesor de la policía científica de Michigan, incorporó un nuevo enfoque metodológico, sentando las bases del método actual. Tosi consideró que no se podía tener en cuenta una sola opción de análisis, puesto que todas ellas eran válidas y complementarias entre sí. Además, incorporó los primeros análisis automáticos para reducir aún más las valoraciones subjetivas realizadas por el experto.

Desde finales del pasado siglo la técnica de la identificación forense de locutores está completamente consolidada: en 1999 expertos en Acústica Forense de numerosos laboratorios europeos celebraron varias reuniones en el seno de la red forense ENFSI. Consecuencia de estas reuniones fue el reconocimiento de los "métodos combinados" como la mejor alternativa de trabajo. La definición de "método combinado" quedó recogida de la siguiente forma: "se consideran métodos combinados de identificación forense de locutores, aquellos que entre sus aproximaciones de estudio incluyen, al menos, el enfoque perceptivo-auditivo, el análisis acústico (sonográfico, oscilográfico, espectrográfico) y el análisis fonético-lingüístico" [2]. A estos análisis puede incorporarse el uso de sistemas de Reconocimiento Automático de Locutores en los casos en los que las condiciones de las muestras lo permitan.

En definitiva, abordar el estudio de una referencia biométrica con un carácter tan variable como es el habla exige un enfoque multidisciplinar que desemboca en la necesaria utilización de una metodología basada en diferentes perspectivas de estudio.

- **Análisis de percepción auditiva:** la memoria auditiva, a medio o largo plazo, nos permite reconocer la voz de las personas con las que nos relacionamos. Los expertos forenses hacen uso de esta facultad para, a partir de la reproducción de pasajes sonoros registrados, centrar su atención sobre determinadas características de los actos de habla de un locutor como el timbre, el acento, las pausas, la tensión articulatoria, la velocidad de elocución…

- **Análisis acústico-espectrográfico:** las peculiaridades detectadas a nivel auditivo se contrastan con las diferentes formas de representación gráfica de la señal de voz (oscilograma, sonograma, espectro), incorporando un enfoque técnico que ayuda al experto a confirmar lo percibido. La forma

de representación que permite obtener información más relevante es el sonograma o espectrograma que aporta una imagen de la distribución de la energía del habla en función de la frecuencia y el tiempo, permitiendo así mismo la observación de las características acústicas de los diferentes sonidos.

- **Análisis fonoarticulatorio/lingüístico:** el análisis fonoarticulatorio pone en relación las peculiaridades observadas a través de los otros análisis con las referencias estándar de articulación de la lengua española. Se centra en la fase de producción del habla, detectando lugar y modo de articulación, grados de tensión-relajación, imprecisiones articulatorias, posibles patologías, etc. El análisis lingüístico abarca los diferentes niveles de la lengua: morfosintáctico, léxico y semántico. Se consideran, por ejemplo, el uso de recursos retóricos, la forma de construcción del discurso, el vocabulario, las influencias dialectales, el estilo de habla, etc.

- **Análisis mediante un Sistema de Reconocimiento Automático:** este análisis, que desarrollamos en el siguiente epígrafe, por ser el utilizado en el presente estudio, se posiciona como una herramienta de gran utilidad para el experto forense al ofrecer una óptica cuantitativa y permitir el análisis con discursos en lengua extranjera.

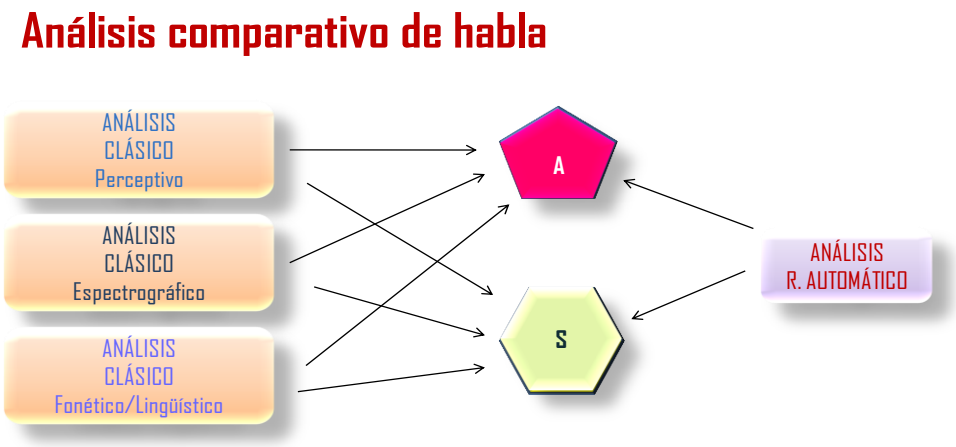

Figura 11. Esquema de "método combinado" de identificación forense de locutores donde "A" es una muestra de voz cuya autoría se conoce y "S" es una muestra de voz de un sospechoso. El estudio de ambas muestras se aborda desde los distintos análisis.

3.3. Sistema de Reconocimiento Automático de Locutores (SRAL).

Un Sistema de Reconocimiento Automático de Locutores es una aplicación informática que utiliza la biometría de voz para realizar la identificación o verificación de la identidad de una persona a partir de sus emisiones habladas.

Los primeros sistemas de Reconocimiento Automático de Locutores aparecen en los años 70, en Estados Unidos. Inicialmente, los laboratorios Bell emplearon coeficientes cepstrum y coeficientes de predicción linear (LPC) para extraer los datos de las señales acústicas emitidas por hablantes [7]. Durante esta década surgen numerosos estudios orientados a la búsqueda de las referencias del habla más idóneas para su tratamiento mediante modelos matemáticos.

En las siguientes décadas se crean diferentes generaciones de sistemas basados en distintas alternativas de parametrización y modelado para perfeccionar la técnica. En la actualidad estos sistemas incorporaran Inteligencia Artificial (IA) basada en el uso de redes neuronales.

En general, un Sistema de Reconocimiento Automático de Locutores realiza de manera autónoma los siguientes procesos:

- **Parametrización**: Extracción de características de la señal (emisión hablada).
- **Modelado** de estas características para obtener un "modelo" de la voz del locutor.
- **Comparación** entre los audios objeto de análisis.
- Obtención de **puntuaciones** de similitud.

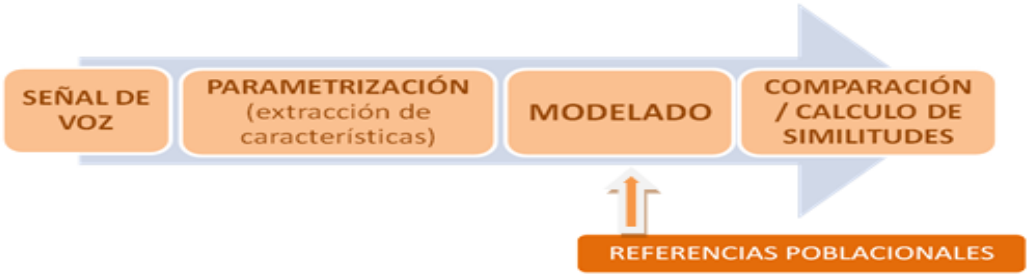

Figura 12. Esquema general de un Sistema de Reconocimiento Automático de Locutores.

En cada una de estas fases existen distintas opciones desde el punto de vista tecnológico y matemático que darán lugar a diferentes tipos de sistemas de reconocimiento automático.

De la señal de voz se puede extraer información en dos niveles: bajo nivel o de parámetros acústicos (nivel fonético) y alto nivel (nivel morfosintáctico, léxico-semántico, prosódico…). La extracción y modelado de características de bajo nivel es más sencilla y da mejores resultados cuando se dispone de grabaciones de corta duración por lo que los sistemas que usan este tipo de parámetros son los más útiles en el entorno forense donde, a menudo, no se dispone de grandes cantidades de habla procedentes de un mismo acto y las muestras suelen presentar alguna degradación. Además, en estos sistemas existen dos opciones de trabajo: con dependencia y con independencia de texto. Los sistemas independientes de texto permiten al experto trabajar con lenguas que no conoce.

Por tanto, un Sistema de Reconocimiento Automático que trabaje con parámetros acústicos o de bajo nivel y que sea independiente de texto será la opción más idónea para trabajar en el entorno forense.

Es el caso de BATVOX, lanzado al mercado en el año 2004 por la empresa española Agnitio, utilizado en numerosos laboratorios de Acústica Forense de todo el mundo y sometido a las evaluaciones comparativas organizadas por el National Institute of Standards and Technology de los Estados Unidos. Este sistema, utilizado

por la sección de Acústica Forense de la Comisaría General de Policía Científica, es el empleado en la realización de los ensayos sobre los que versa el presente trabajo.

A grandes rasgos y siguiendo el esquema general mostrado anteriormente, el funcionamiento de este sistema es el siguiente:

1. Parte de una **locución** procedente de un locutor no clasificado.

2. En la fase de **parametrización**, convierte la señal acústica de entrada en una serie de vectores de características que extraigan de forma eficiente la información de locutor presente en la señal de voz. Para ello utiliza lo que se conoce como análisis localizado: mediante un enventanado divide la señal en intervalos temporales del orden de milisegundos solapados entre sí, ya que la señal a largo plazo presenta mucha variabilidad; sin embargo, a corto plazo (5-10 ms) es una onda quasi-estacionaria y pseudoperiódica.

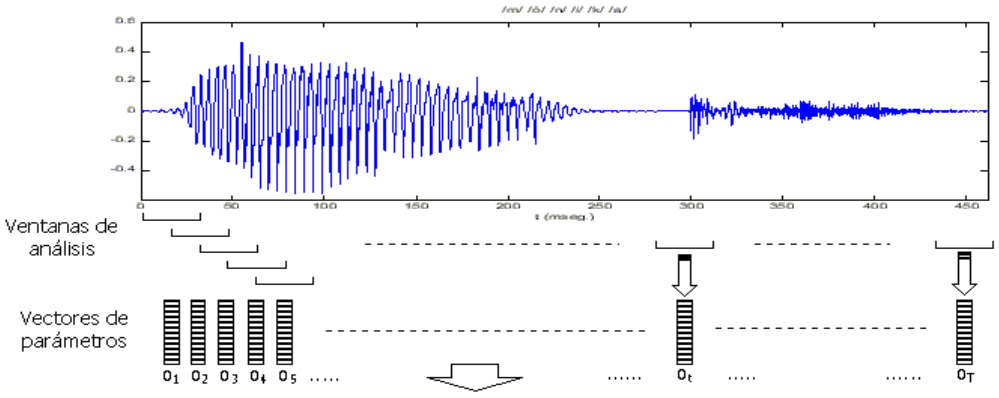

Figura 13. Esquema de enventanado de señal y obtención de vectores de características.

De esta manera, la señal inicial se convierte en un conjunto de vectores de coeficientes que recogen las características de la muestra a nivel acústico. El número de vectores de características que se obtiene dependerá de la longitud del audio.

Antes de realizar el siguiente paso, el sistema reduce el número de datos con el que tendrá que trabajar para obtener el modelo utilizando dos estrategias:

a) Se queda con la información relevante y con capacidad discriminativa, la procedente del tracto vocal que contiene las características de la estructura de resonancia, desechando aquella información que se genera a nivel glotal, referente a frecuencia fundamental.

b) Imita el funcionamiento del oído humano utilizando como escala de frecuencias la Escala MEL (subjetiva).

3. El **modelado** de la señal es la clave del funcionamiento del sistema. En esta fase se tratan los datos obtenidos en la fase anterior para que el archivo obtenido reúna las características y propiedades específicas del locutor y sea de un tamaño manejable por el sistema.

En nuestro caso se utiliza un modelado de mezcla de 1024 gaussianas (GMM) que permite dimensionar adecuadamente el modelo de manera que, sea cual sea la longitud (duración) inicial del audio, el resultado sea un archivo de dimensiones fijas. Así el número de datos para procesar no dependerá de la duración de la grabación.

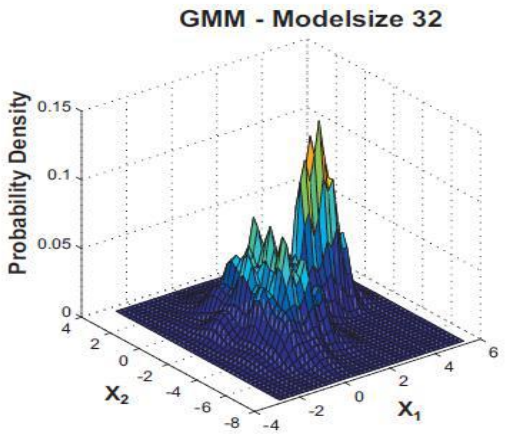

Figura 14. Ejemplo de modelado de mezcla de gaussianas (32).

Los modelos de locutor obtenidos son entrenados mediante un procedimiento de adaptación a una UBM (Universal Background Model), que es un modelo representativo de la generalidad del habla humana generado a partir de una gran cantidad de habla de un gran número de personas de ambos géneros en distintas

circunstancias. De esta forma, se consigue extraer las características más relevantes del audio modelado, que más lo distinguirán de las típicas del ser humano. En concreto, Batvox posee dos modelos universales, uno de voz femenina y otro de voz masculina. Cada uno de los fondos documentales consta de más de mil horas de habla en su género respectivo, proveniente de distintas lenguas maternas, tanto de habla conversacional como lectura, registradas en diferentes soportes y canales (analógico, digital, microfónico, telefónico, etc.).

4. En la fase de **comparación**, una vez obtenidos los vectores de características correspondientes a la señal de voz de entrada, y teniendo ya un modelo del locutor objeto de estudio, realiza el cotejo entre los dos tipos de muestras.

En esta etapa es fundamental el uso que el sistema hace de las poblaciones de referencia de que dispone. Una población de referencia es un conjunto de modelos de locutor con unas características determinadas en cuanto a las variables que el sistema tiene en cuenta para hacer las comparaciones (género del locutor, canal de transmisión y grabación del audio, longitud del mismo, tipo de habla, idioma). Las características del modelo de locutor analizado deben estar en consonancia con las de los modelos de la población de referencia para que la comparación sea robusta y los resultados obtenidos, fiables.

En cualquier tipo de comparación los elementos objeto de cotejo se parecerán más o menos entre sí en función de las referencias que tomemos para observar las similitudes y disimilitudes. Como ejemplo gráfico para ilustrar este concepto: las similitudes que observamos entre dos personas de una misma procedencia geográfica (tez oscura, labios gruesos, nariz ancha, ojos y pelo negros en personas de África central) son mayores si tomamos como referencia para establecer la comparación una población de personas de otra zona (tez clara, pelo rubio, ojos claros, en personas del norte de Europa), es decir, se nos parecerán más de lo que en realidad se parecen. Sin embargo, las diferencias o similitudes que apreciaremos serán más consistentes y fundamentadas si tomamos como referencia una población de personas de la misma procedencia.

Cuanto más amplio sea el abanico de poblaciones de referencia disponibles mejor se podrá optimizar el trabajo realizado; en este sentido, el experto selecciona las características de la población que considera más adecuada al caso y el propio sistema elige, entre todos los modelos disponibles de esa población, aquellos que mejor se adaptan al locutor objeto de estudio (con 35 modelos el rendimiento es óptimo [7]).

5. Las **puntuaciones de similitud** vienen expresadas en forma de relaciones de verosimilitud (LRs) a través del cociente de dos probabilidades: la probabilidad de que, perteneciendo la voz dubitada a un hablante determinado, se produzca la evidencia E (factor de similitud) y la probabilidad de que dada la hipótesis complementaria (la voz dubitada no pertenece a ese hablante), se produzca la evidencia E (factor de tipicidad). Por tanto, el valor del LR será mayor que uno si lo más probable es que la voz dubitada pertenezca al locutor indubitado. Si por el contrario es más probable que la voz dubitada no pertenezca a éste, el valor del LR será menor que uno.

Hemos explicado las tareas que el sistema que nos ocupa realiza de manera autónoma: extracción de las características de las señales de voz, entrenamiento con una UBM para obtener un modelo de la voz indubitada del sospechoso, comparación entre las grabaciones utilizando poblaciones de referencia y obtención de puntuaciones de similitud (LR).

Sin embargo, en la realización de un cotejo de voces mediante este SRAL, la interacción del experto forense es pieza clave para que los resultados obtenidos, reflejados en una conclusión final, tengan la consistencia que requiere la emisión de un informe sobre análisis comparativo de voces en el marco de un proceso penal. Esta intervención se produce tanto en la fase previa de preparación de las muestras, donde es necesario determinar si estas son aptas o no para el análisis y editarlas adecuadamente para que no aparezca en un mismo audio habla de distintos locutores, como en la fase de selección de la población de referencia más ajustada posible y en la toma final de decisiones, interpretando los resultados aportados de manera numérica por el sistema en función de todas las variables que existan en el caso.

4. Voz sintética: síntesis de voz. Sistemas TTS. Inteligencia artificial. Clonación de voz.

4.1. Síntesis de voz. Métodos más utilizados.

La voz es el principal instrumento de una de las necesidades básicas del ser humano, comunicarse con sus semejantes. En un mundo en el que las "máquinas" cobran cada vez mayor protagonismo, adquiere especial importancia dotarlas de la facultad de comunicar y de hacerlo de manera similar a como lo hacemos las personas.

Desde que a principios del siglo XX apareciese el primer sintetizador de voz eléctrico, la Tecnología del Habla se ha esforzado en crear sistemas capaces de generar voz artificial que, más allá de resultar inteligible, sea lo más parecida posible a la voz humana natural.

La síntesis de voz posibilita que la prestación de determinados servicios vaya acompañada de la comunicación oral, lo que resulta especialmente interesante en el caso de personas con limitaciones físicas (p.e. con deficiencias visuales o de producción del habla). Todos recordamos a Stephen Hawking haciendo uso de un sintetizador de voz para comunicarse [8]. Este astrofísico, cuyos estudios fueron relevantes en el conocimiento de los agujeros negros, padecía Esclerosis Lateral Amiotrófica, pero su enfermedad no le impidió seguir investigando y trasladando su conocimiento al mundo científico gracias a la ayuda de este sistema, con el que componía palabras y frases usando la contracción voluntaria de unas de sus mejillas.

Los métodos más comúnmente utilizados para producir habla de manera artificial (síntesis de formantes, síntesis articulatoria y síntesis concatenativa), que desarrollaremos brevemente más adelante, se han visto actualmente superados con la entrada en escena de la inteligencia artificial. El uso de redes neuronales para generar habla permite reproducir cualquier discurso a partir de tan solo unos segundos de grabación de voz natural y con unas características tan similares a la original que a veces resulta casi imposible distinguirlas. Así, por ejemplo, podemos escuchar y ver al expresidente de EEUU Barak Obama realizando un discurso que nunca hizo [9]; o escuchar la voz de Francisco Franco recitando la letra de una conocida canción escrita años después de su muerte [10,11].

Como ya hemos puntualizado, el objetivo de los métodos utilizados para generar voz es conseguir sonidos que se asemejen lo más posible al habla humana, tanto en la propia realización de los sonidos como en la concatenación de los mismos para crear palabras y frases, introduciendo las debidas pausas y con una prosodia adecuada. Los bloques de síntesis de las aplicaciones que generan voz pueden utilizar distintos métodos [12-14]:

1. **Síntesis de formantes**: Este método es el más sencillo, puesto que utiliza un modelo acústico creado previamente que dota a la voz sintetizada de un carácter robótico, aunque perfectamente inteligible. Se basa en un diseño muy simplificado del aparato fonador, asociando a cada alófono (realización de un sonido) un conjunto de valores de frecuencia central y ancho de banda de sus formantes teóricos. Fue muy utilizado en los primeros videojuegos y a día de hoy lo podemos encontrar en muchos lectores de pantalla. Al trabajar de forma sencilla, sin tener que recurrir a bases de datos de voz humana, a diferencia de los siguientes métodos, permite sintetizar voz a gran velocidad, lo cual resulta muy útil para personas con discapacidad visual. Además, se puede utilizar en sistemas embebidos donde la memoria y el procesador son limitados, por ejemplo, una máquina expendedora.

2. **Síntesis articulatoria**: Se sustenta en un modelo de las características físicas del tracto vocal (longitud y aéreas transversales) que incluye tanto la información relativa a la vibración de las cuerdas vocales como la de los diferentes estados del sistema tracto vocal [15]. La desventaja de este método es que requiere controlar una gran cantidad de información sobre la posición y movimiento de los órganos articulatorios y su coordinación.

3. **Síntesis concatenativa**: Se basa en la unión de pequeños segmentos de voz grabados, por lo que necesitan de una base de datos de muestras de habla humana. Esto resulta en una voz más natural, de mayor calidad que los anteriores sistemas, aunque se sigan percibiendo discontinuidades en el discurso. Existen tres métodos para realizar esta síntesis:

- Síntesis de selección de unidades: Cada acto del habla registrado ha sido segmentado en fonemas, sílabas, palabras, frases y oraciones con ayuda de un Sistema de Reconocimiento del Habla. Esta operación se realiza multitud de veces, para poder disponer de repeticiones de la misma unidad. Como no utiliza un modelo de producción de voz, las unidades no están parametrizadas, y esto repercute en el tamaño de la base de datos, lo que la hace poco práctica. A su vez, como las unidades a unir se han registrado en contextos diferentes, están coarticuladas por los fonemas vecinos en el momento de la grabación. Esto produce una discontinuidad en cada unión. Se han realizado estudios de mejora en ambos sentidos, sin llegar a ser del todo satisfactorios.

- Síntesis por difonemas: Comprenden parte de dos fonemas, centrándose en las transiciones entre sonidos que se dan en el lenguaje hablado y los utilizan como unidades básicas en la síntesis, lo que resulta en transiciones suaves entre segmentos. En castellano existen unos 800 difonemas. Este método fue creado para hacer más manejable la base de datos, aunque actualmente su uso es prácticamente nulo debido a que genera una voz muy robótica.

- Síntesis de dominio específico: Este método une palabras y frases previamente registradas para crear enunciados completos. Como indica su nombre, su uso está focalizado a un ámbito en concreto como puede ser anuncios o avisos de cualquier tipo, por ejemplo, los disponibles en los relojes inteligentes o *smartwatch*. Al requerir una base de datos mucho más restringida en contenido, la voz sintetizada se llega a percibir mucho más natural que las anteriores al ser semejante en prosodia y entonación que la grabación

original. Por el contrario, este método no puede ser utilizado para otros fines.

4.2. Conversores texto a voz (TTS).

Uno de los usos más comunes de la síntesis de voz es el de los conversores texto a voz (CTV), en inglés: *Text To Speech* (TTS). La conversión texto-voz "es la generación, por medios automáticos, de la secuencia de sonidos que produciría una persona al leer un texto cualquiera en voz alta" [16].

El objetivo de este tipo de aplicaciones es construir de manera completamente autónoma, a partir de un texto escrito arbitrario, un discurso similar al de un ser humano y dotarlo de la mayor naturalidad posible.

El proceso de lectura de texto es bastante complejo en su conjunto, aunque de forma genérica se puede presentar como un sistema secuencial dividido en tres bloques principales: proceso lingüístico, prosodia y síntesis, tal y como se muestra en la siguiente imagen:

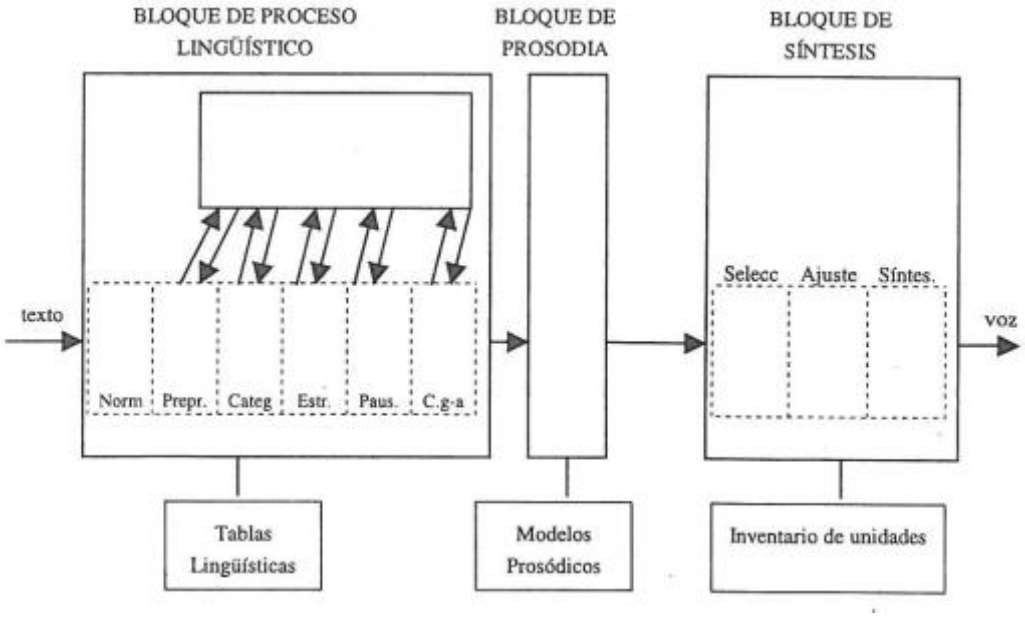

Figura 15. Esquema de un sistema de conversión texto-voz con estructura secuencial.

1. Bloque de proceso lingüístico: para poder leer un texto, el sistema tiene que conocer qué sonidos existen y cómo producirlos. Tras realizar una serie de etapas, que se nombran seguidamente, se consigue obtener una cadena de alófonos, sonido propio de la pronunciación de un fonema que puede variar según su posición en la palabra o en la sílaba y en relación con los sonidos vecinos.

2. Bloque generador de parámetros prosódicos: la prosodia, en el ámbito de la tecnología del habla, se refiere principalmente a frecuencia fundamental (entonación), duración de los alófonos (cantidad) y energía (intensidad), esta última menos relevante. En este tipo de información se manifiestan tanto elementos lingüísticos (interrogación, pausas, acentos, etc.) como no lingüísticos, un estado de ánimo, por ejemplo. El sistema solo podrá controlar los lingüísticos, aunque se pueden aplicar modelos específicos para situaciones concretas con el fin de acercarnos al habla natural.

3. Bloque de síntesis de voz. La secuencia de alófonos que se debe generar y los datos referentes a la prosodia son el punto de partida del trabajo de este bloque y constituyen el núcleo del sintetizador, que aplicará alguno de los métodos ya descritos: síntesis de formantes, síntesis articulatoria o síntesis concatenativa.

4.3. Inteligencia artificial. Redes neuronales artificiales.

El aumento de la capacidad de los ordenadores para realizar procesos complejos ha propiciado la evolución de las ciencias de la computación y la aparición de sistemas capaces de imitar la inteligencia del ser humano utilizando algoritmos que permiten analizar y clasificar grandes cantidades de datos, procesarlos y obtener información suficiente para tomar decisiones. Los modelos de inteligencia artificial basan su eficacia en su capacidad para aprender de forma automática. Este aprendizaje

autónomo, conocido como *machine learning*, ha encontrado el mejor caldo de cultivo en un mundo en el que la información se mide en miles de millones de datos.

Uno de los modelos de inteligencia artificial con mayor capacidad de aprendizaje es el de redes neuronales, algoritmos existentes ya en los años 60 que no han podido ser utilizados hasta décadas después debido a las limitaciones impuestas por la tecnología [17].

Una red neuronal artificial es un algoritmo matemático que imita el comportamiento del sistema nervioso y del cerebro humano, creando un procedimiento de interconexión entre capas de neuronas similar a los procesos de sinapsis de las neuronas del cerebro [18-20]. Estas redes se organizan en nodos o elementos de proceso (neuronas) agrupados formando capas y sus enlaces (conexiones entre los nodos) del modo que se muestra esquemáticamente en la figura 16.

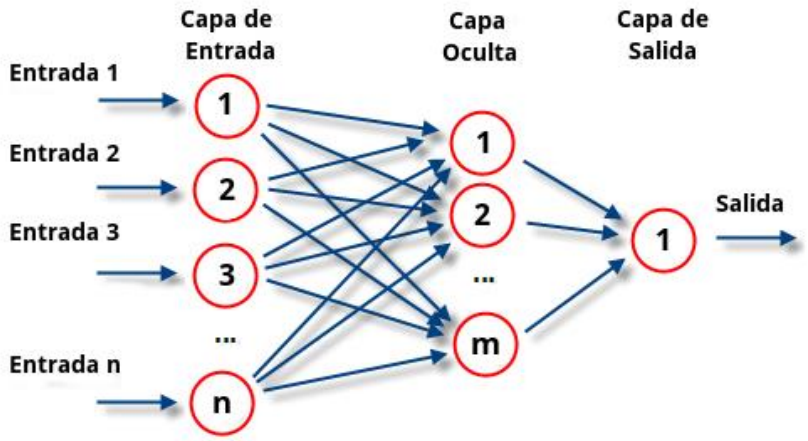

Figura 16. Esquema que representa la estructura básica de una red neuronal artificial donde la capa de entrada es la que recibe los datos, la capa oculta es la capa intermedia que puede ser múltiple, y la capa de salida, que recibe los datos de las anteriores y da una respuesta.

La red recibe una información de entrada en forma de valores de determinadas variables (p.e. el precio de un producto, la fecha de compra, el número de unidades adquiridas, etc.). Esta información es procesada por cada una de las neuronas artificiales de la capa de entrada, que va identificando patrones, dando lugar a una

información de salida, que se transmite a las neuronas de la capa siguiente, continuando su recorrido por el resto de la red con el fin de obtener predicciones [21-23]. En esta transmisión interviene una "función de activación" cuya misión es seleccionar la manera en la que debe salir la información hacia la siguiente capa. Por ejemplo, nos interesa usar una "función identidad" si queremos que la información de salida no se modifique, o una "función escalón" si queremos filtrar determinados valores [24].

Cada nodo neuronal tiene un peso asignado en función del cual modifica el input recibido de manera que la información procesada consiste en una combinación del producto de las entradas por sus pesos. Para que la red neuronal realice correctamente la función esperada es necesario someterla a un proceso de entrenamiento en el que se modifican los pesos para conseguir los valores de salida adecuados [25,26]. Esta es la forma en la que la red "aprende" y es posible obtener resultados muy correctos incluso cuando los datos de los que se dispone son muy diferentes a los que se han usado para el entrenamiento.

El uso de la inteligencia artificial basada en redes neuronales es muy amplio, es útil en sectores como la industria, la economía o la medicina [27]. Su incorporación a las tecnologías del habla ha resultado en lo que conocemos como clonación de voz: a partir de tan sólo unos segundos de habla es posible convertir cualquier texto escrito en un discurso que se percibirá muy parecido a la voz original, reproduciendo no sólo su timbre sino los rasgos suprasegmentales (características de entonación y acento), el estilo de habla y la prosodia, con el consiguiente aumento de la naturalidad y la expresividad. [28].

Uno de los sistemas más conocidos de clonación de voz es la plataforma Lyrebird que, con solo crear una cuenta, permite al usuario generar su propia voz artificial. Actualmente, todas las grandes multinacionales de la comunicación disponen de sistemas de este tipo: en 2017 Google presentó su proyecto Tacotrón 2; Amazon Polly, el generador de voz sintética de la empresa del mismo nombre, ofrece voces en distintos idiomas y estilos de habla; IBM lanzó en 2018 su asistente de voz Watson [29,30], etc.

Actualmente es posible encontrar en la red numerosos sistemas que disponen de diferentes voces generadas mediante el uso de redes neuronales, en distintos idiomas y dialectos y en ambos géneros para transformar textos escritos en voz. Varios de estos sistemas, disponibles de manera gratuita para textos de una duración limitada, han sido utilizados para realizar los ensayos que se exponen en la parte experimental de este trabajo

5. Parte experimental. Análisis comparativo con voz sintética mediante SRAL.

5.1. Descripción de los elementos del sistema y flujo de trabajo.

El trabajo del Sistema de Reconocimiento Automático de Locutores BATVOX 4.1 se vertebra en torno a lo que se denomina "caso", como unidad fundamental. En cada caso utiliza archivos de audio que clasifica, según el carácter de los mismos y las tareas en las que se emplean, en los tipos siguientes:

- AUDIOS TEST: aquellas grabaciones de voz atribuida de las que se desconoce si pertenecen al sospechoso (voz de carácter dubitado).
- AUDIOS DE ENTRENAMIENTO (MODELO): grabaciones que registran la voz del sospechoso y se usan para generar un modelo de locutor (carácter indubitado).
- AUDIOS DE IMPOSTOR: grabaciones de voz que tienen las mismas características de registro que los audios test y se sabe, a ciencia cierta, que no pertenecen al sospechoso.

Todos estos archivos deben reunir unos requisitos en cuanto a **formato de audio, duración de la locución y relación señal-ruido (SNR)**. Así, el sistema establece que el formato de audio debe ser wav con muestreado a una frecuencia de 8000 Hz, con 16 bits de cuantificación y en un solo canal. En cuanto a la duración de las locuciones, en los audios test el habla neta y continua de cada locutor (procedente del mismo acto de grabación) deberá tener una longitud mínima de 7 segundos y preferiblemente superior a 15 segundos. En el caso de los audios utilizados para generar un modelo de locutor

(audios de entrenamiento) la duración deberá ser de 30 segundos como mínimo y preferiblemente superior a 60 segundos. Respecto a la calidad sonora de las grabaciones el sistema recomienda no utilizar registros con una relación señal-ruido (SNR) inferior a 10 dB, considerándose óptimo un valor superior a 15 dB.

Para realizar las tareas comparativas con suficiente fiabilidad es necesario tener en cuenta ciertas características referidas a la naturaleza de las grabaciones. En este sentido, el SRAL, considera las siguientes variables:

- **Sexo** del hablante (hombre, mujer).
- **Canal** de transmisión (microfónico, telefónico, GSM (Sistema Global de comunicaciones Móviles), vídeo, …) y grabación (digital, cinta magnetofónica, …).
- **Idioma**.
- **Tipo de habla** (texto leído, espontánea, conversacional).

Como se explicó anteriormente, las comparaciones se realizan tomando como referencia poblaciones de locutores formadas por grabaciones cuyas características en cuanto a las variables mencionadas estén en consonancia con las del modelo de sospechoso. Por ejemplo, si nuestra grabación indubitada (modelo) pertenece a un varón que lee un texto en lengua inglesa, a través de un micrófono y se ha registrado en un soporte digital deberemos seleccionar, de entre las poblaciones de referencia disponibles, una que esté formada por grabaciones de varones, en lengua inglesa, con texto leído y en soporte digital. Será necesario disponer de una población representativa del modelo formado por, al menos, 25 modelos de similares características.

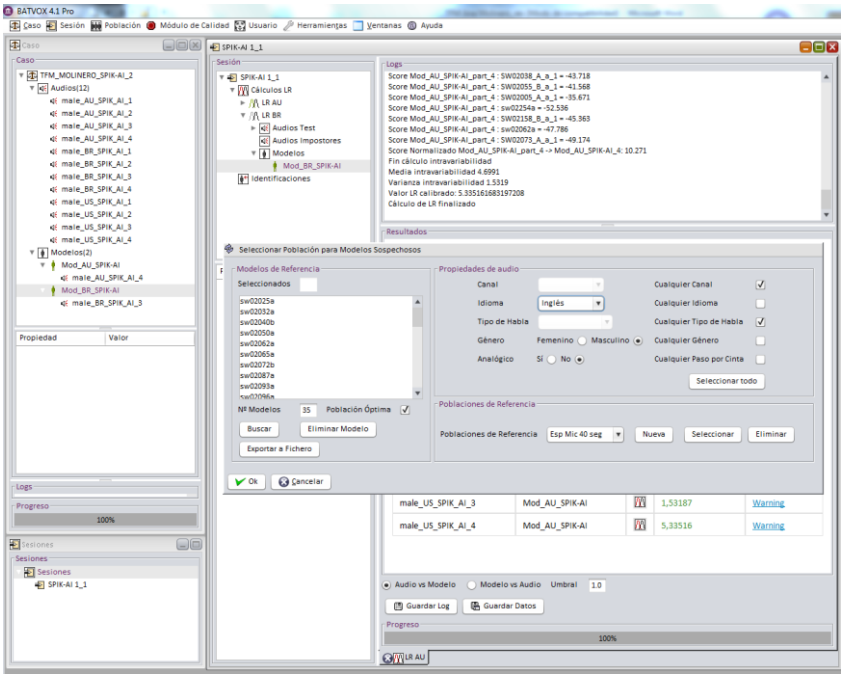

Figura 17. Captura de pantalla de la interfaz del software BATVOX 4.1 en la que se muestra un ejemplo de caso con la selección de una población de referencia para un modelo de sospechoso.

Hay que señalar la importancia de la elección de una población de referencia bien ajustada ya que dos grabaciones con las mismas características de este tipo (sexo, canal, idioma, tipo de habla) pueden llegar a puntuar más alto entre sí siendo de personas distintas que dos grabaciones de la misma persona que tengan distintas características, lo que daría lugar a resultados incorrectos.

En este punto juega un papel fundamental la experiencia del experto y su conocimiento del funcionamiento del sistema ya que, si bien lo deseable es lo anterior, también es cierto que **no es absolutamente necesario un ajuste perfecto para obtener buenos resultados, siempre y cuando se conozca la forma en la que los desajustes afectan al rendimiento del sistema.**

La estructura básica de un caso práctico de comparación realizada con BATVOX sería la siguiente:

Figura 18. Estructura básica de un caso práctico de comparación realizada con BATVOX.

En el planteamiento del caso se establece que el sistema compare los audios test y los audios de impostor con el modelo de sospechoso. Introducir los audios de impostor en las tareas comparativas permitirá determinar si el sistema es capaz de discriminar correctamente ya que conocemos a priori que estos audios no pertenecen al sospechoso. Por tanto, si los valores de los LR relativos a los audios de impostor son menores que uno sabremos que la capacidad de discriminación es buena. Esto nos afianzará en los valores de LR que obtengamos para los audios test.

Siguiendo esta estructura básica hemos diseñado las distintas comparaciones entre muestras que se realizan en este trabajo.

5.2. Material objeto del análisis.

Para la obtención de los archivos de audio de voz artificial se han utilizado tres aplicaciones gratuitas de conversión texto a voz que disponen de galerías con diferentes voces generadas mediante el uso de redes neuronales.

Las aplicaciones elegidas son: SPIK-AI, NUANCE VOCALIZER y Play HT, todas ellas con la opción de conversión online disponible [31-33].

Una vez testeadas dichas aplicaciones, hemos comprobado que el contexto en el que ofrecen una mayor variedad de voces es el caso de **voz de varón en lengua inglesa**, por lo que este ha sido el escenario elegido. Así, a partir de varios textos

escritos en lengua inglesa y usando las versiones online de estas aplicaciones se han generado los archivos objeto de estudio.

SPIK-AI nos ofrece tres voces de varón en lengua inglesa de distintos dialectos (inglés británico, inglés de EEUU e inglés de Australia) y se obtienen cuatro archivos de audio para cada locutor. El resultado de la conversión se descarga en formato mp3, formato de audio con compresión con pérdidas que, si bien no es lo más recomendable, se comprueba que presenta unas características de compresión compatibles con las exigencias del SRAL. Este extremo deberá tenerse en cuenta a la hora de interpretar los resultados.

En la tabla 1 se relacionan los archivos obtenidos mediante esta aplicación, su duración y el valor de la relación señal-ruido, calculados por el sistema.

ARCHIVO DE AUDIO	DURACIÓN (s)	SNR (dB)
male_AU_SPIK_AI_1	27,77	25,619
male_AU_SPIK_AI_2	15,84	24,661
male_AU_SPIK_AI_3	23,02	26,861
male_AU_SPIK_AI_4	44,28	25,237
male_BR_SPIK_AI_1	18,60	29,243
male_BR_SPIK_AI_2	14,56	20,163
male_BR_SPIK_AI_3	58,00	27,117
male_BR_SPIK_AI_4	29,75	24,221
male_US_SPIK_AI_1	15,54	22,219
male_US_SPIK_AI_2	27,47	29,735
male_US_SPIK_AI_3	31,95	27,216
male_US_SPIK_AI_4	59,60	26,849

Tabla 1. Archivos generados con SPIK-AI en los dialectos australiano (AU), británico (BR) y americano (US).

NUANCE VOCALIZER presenta voces en inglés en las mismas variedades dialectales que el anterior (inglés británico, ingles de EEUU e inglés de Australia). Se dispone

de seis locutores distintos (Daniel, Lee, Evan, Malcolm, Nathan y Tom) y se obtienen cuatro archivos de audio para cada uno, relacionados en la tabla 2. Estos archivos se descargan en formato wav [48.000 Hz, 16 bit, mono].

ARCHIVO DE AUDIO	DURACIÓN (s)	SNR (dB)
male_BR_DANIEL_NV_1	21,181	26,371
male_BR_DANIEL_NV_2	43,6	20,969
male_BR_DANIEL_NV_3	15,49	19,379
male_BR_DANIEL_NV_4	12,79	25,237
male_AU_LEE_NV_1	22,24	29,86
male_AU_LEE_NV_2	42,94	23,639
male_AU_LEE_NV_3	16,091	26,407
male_AU_LEE_NV_4	14,19	27,38
male_US_EVAN_NV_1	19,71	28,484
male_US_EVAN_NV_2	41,32	26,477
male_US_EVAN_NV_3	14,66	24,627
male_US_EVAN_NV_4	12,461	29,416
male_BR_MALCOLM_NV_1	21,83	26,229
male_BR_MALCOLM_NV_2	44,521	18,11
male_BR_MALCOLM_NV_3	16,191	19,002
male_BR_MALCOLM_NV_4	13,721	29,591
male_US_NATHAN_NV_1	20,25	28,718
male_US_NATHAN_NV_2	43,081	22,755
male_US_NATHAN_NV_3	15,471	21,504
male_US_NATHAN_NV_4	12,83	28,221
male_US_TOM_NV_1	20,8	26,401
male_US_TOM_NV_2	43,07	20,98
male_US_TOM_NV_3	14,931	21,865
male_US_TOM_NV_4	12,29	28,053

Tabla 2. Archivos generados con NUANCE VOCALIZER.

Por último, la aplicación **Play HT** ofrece grabaciones de voz de varón en inglés de EEUU con seis voces diferentes (Guy, James, Joseph, Mark, Noah y Richard). Los archivos se han descargado en formato wav [44.100 Hz, 16 bit, mono] y presentan las características que se muestran en la tabla 3.

ARCHIVO DE AUDIO	DURACIÓN (s)	SNR (dB)
male_US_GUY_PHT	36,291	29,273
male_US_JAMES_PHT	17,191	25,471
male_US_JOSEPH_PHT	17,710	26,781
male_US_MARK_PHT_1	35,151	26,624
male_US_MARK_PHT_2	16,771	29,939
male_US_NOAH_PHT_1	32,620	28,564
male_US_NOAH_PHT_2	16,980	30,000
male_US_RICHARD_PHT	16,511	21,736

Tabla 3. Archivos generados con PlayHT.

Los archivos que en las tablas aparecen en color azul son aquellos que reúnen las condiciones de duración exigidas por el sistema para ser entrenados como modelos de locutor. Se realizarán, por tanto, en cada caso, las comparaciones de estas grabaciones con el resto de las obtenidas con la misma aplicación.

5.3. Estudio comparativo mediante un SRAL.

Como ya hemos expuesto, está consensuado que el análisis comparativo de voces, dado el carácter variable de esta referencia biométrica, debe llevarse a cabo utilizando distintas aproximaciones de estudio.

En el caso que nos ocupa nos hemos centrado en la observación de la respuesta de un Sistema de Reconocimiento Automático de Locutores ante grabaciones de voz no

natural. Un análisis integral en el que tuviésemos en cuenta otras perspectivas de estudio escapa al alcance de este trabajo.

No obstante, como paso previo, ha sido necesaria una valoración de las muestras a nivel perceptivo y acústico-espectrográfico, al objeto de determinar si existen similitudes y diferencias apreciables entre los distintos locutores.

En una primera aproximación, se observa que el habla presenta continuidad y una coarticulación adecuada que se corresponden en la estructura de la señal acústica con una distribución de la energía similar a la de una voz natural.

Perceptivamente se observan notables disimilitudes entre los locutores en lo referente a la cualidad tímbrica, el tono y la velocidad de elocución. La observación de la representación gráfica de la señal acústica permite establecer diferencias en las estructuras de resonancia, como se muestra en las figuras 19 y 20 en las que pueden apreciarse distintas alturas formánticas en la realización del mismo sonido por dos locutores distintos [7].

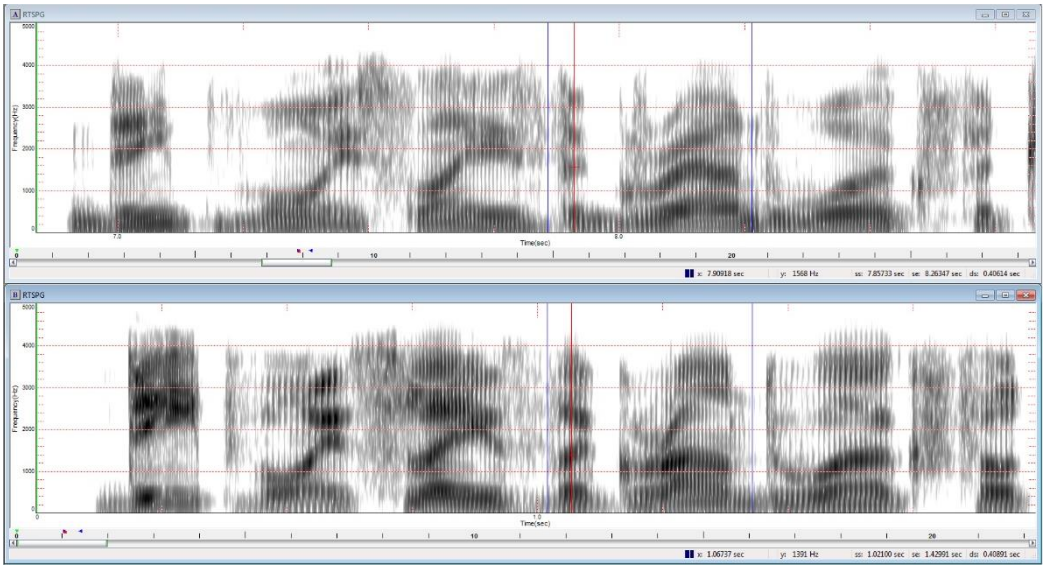

Figura 19. Representaciones espectrográficas correspondientes a la frase "Deep Voice lays the groundwork for…", extraída de los archivos denominados "male_AU_LEE_NV_1" (ventana A) y "male_BR_DANIEL_NV_1" (ventana B).

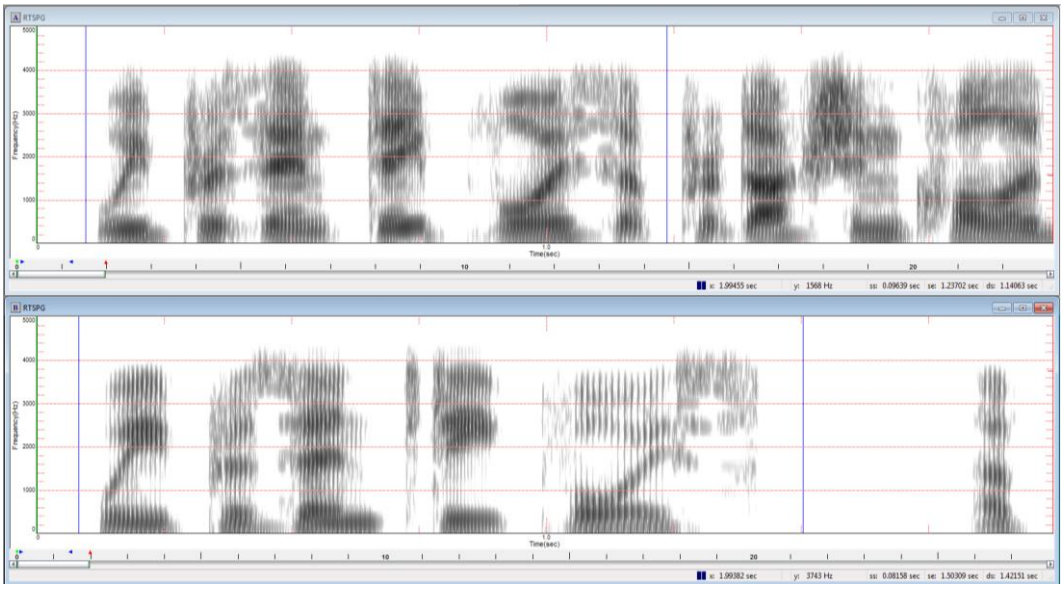

Figura 20. Representaciones espectrográficas correspondientes a la frase "We present Deep Voice…", extraída de los archivos denominados "male_AU_SPIK-AI_1" (ventana A) y "male_BR_SPIK-AI_1" (ventana B).

A continuación, detallamos los análisis realizados para las tres aplicaciones elegidas.

5.3.1. Caso 1. SPIK-AI.

Los audios seleccionados para generar modelos de locutor en este caso son:

ARCHIVO DE AUDIO	MODELO GENERADO
male_AU_SPIK_AI_4	Mod_AU_SPIK-AI
male_BR_SPIK_AI_3	Mod_BR_SPIK-AI
male_US_SPIK_AI_4	Mod_US_SPIK-AI

Tabla 4. Audios seleccionados de SPIK-AI para generar modelos de locutor.

Se lanza la comparación de cada uno de ellos contra el resto de audios obtenidos con esta misma aplicación.

La población de referencia seleccionada consiste en los 35 modelos más competitivos de entre los 71 disponibles en voz de varón, en distintos canales, soporte digital y lengua inglesa. Esta población parece representar adecuadamente al modelo de locutor en el caso de los dos primeros modelos, no así para el modelo de la variedad dialectal de Estados Unidos por lo que no ha sido posible seguir adelante con el cotejo en este último caso.

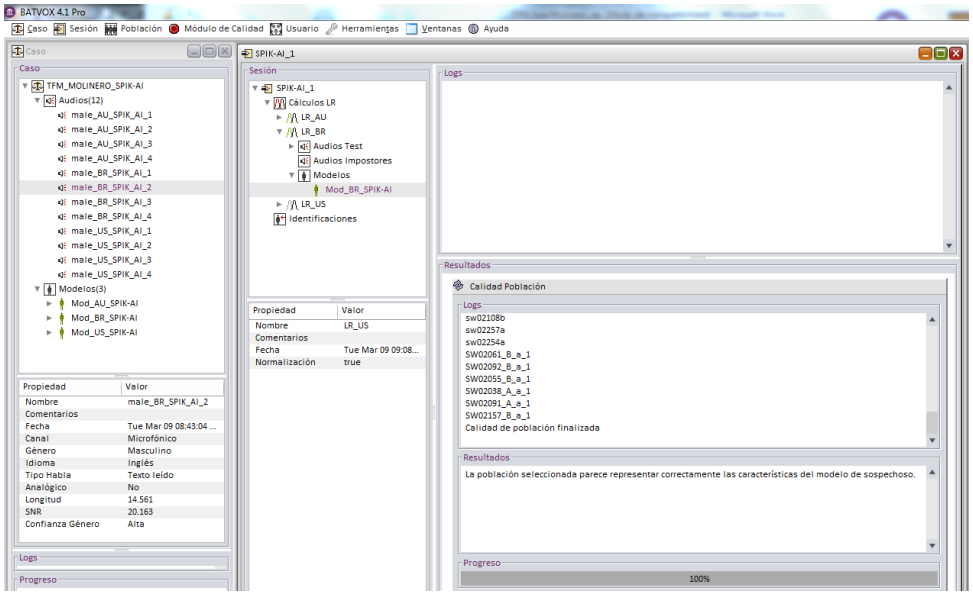

Figura 21. Captura de pantalla de la interfaz del software BATVOX 4.1 en la que se observa que el sistema determina que la población de referencia es representativa del modelo de locutor Mod_BR_SPIK-AI.

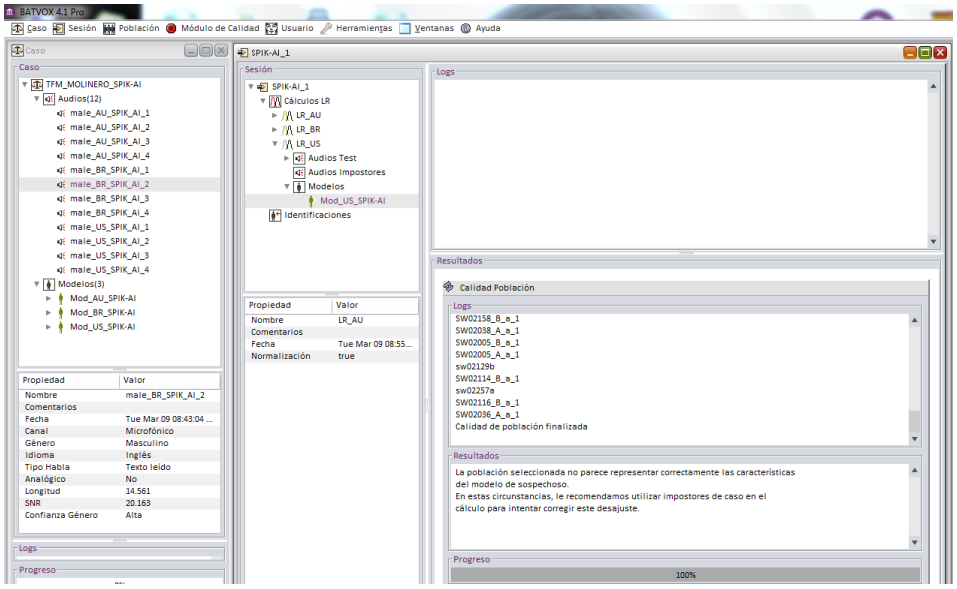

Figura 22. Captura de pantalla de la interfaz del software BATVOX 4.1 en la que se observa que el sistema determina que la población de referencia no es representativa del modelo de locutor Mod_US_SPIK-AI.

Los resultados obtenidos (expresados en valores de LR) se muestran en las tablas siguientes:

	Mod_AU_SPIK-AI
male_AU_SPIK_AI_1	1.00E+10
male_AU_SPIK_AI_3	1.00E+10
male_AU_SPIK_AI_2	1.00E+10
male_BR_SPIK_AI_1	1.68E+08
male_BR_SPIK_AI_2	7.28E+05
male_BR_SPIK_AI_4	1.52E+04
male_BR_SPIK_AI_3	9.95E+03
male_US_SPIK_AI_4	5.34E+00
male_US_SPIK_AI_3	1.53E+00
male_US_SPIK_AI_2	1.28E+00
male_US_SPIK_AI_1	6.70E-02

Tabla 5. Resultados obtenidos de cotejo entre los archivos generados con SPIK-AI y el modelo de locutor Mod_AU_SPIK-AI.

	Mod_BR_SPIK-AI
male_BR_SPIK_AI_1	9.98E+09
male_BR_SPIK_AI_2	9.26E+09
male_AU_SPIK_AI_1	4.41E+09
male_AU_SPIK_AI_4	1.91E+07
male_BR_SPIK_AI_4	3.86E+06
male_AU_SPIK_AI_3	1.05E+06
male_AU_SPIK_AI_2	2.56E+04
male_US_SPIK_AI_4	3.70E+02
male_US_SPIK_AI_2	1.55E+02
male_US_SPIK_AI_3	6.12E+01
male_US_SPIK_AI_1	9.55E+00

Tabla 6. Resultados obtenidos de cotejo entre los archivos generados con SPIK-AI y el modelo de locutor Mod_BR_SPIK-AI.

Como puede observarse, tanto las puntuaciones de los archivos que recogen voz del mismo locutor que la del modelo como el resto son en su mayoría superiores a uno, lo que significa que para el sistema todas las grabaciones pertenecen a un mismo hablante. Sin embargo, en ambos casos, los archivos del mismo locutor aparecen agrupados en la parte superior de la tabla con unas puntuaciones de mayor orden que el resto, principalmente en el caso del hablante del dialecto australiano.

5.3.2. Caso 2. NUANCE VOCALIZER.

Los audios seleccionados para generar modelos de locutor en este caso son:

ARCHIVO DE AUDIO	MODELO GENERADO
male_BR_DANIEL_NV_2	Mod_DANIEL_NV
male_AU_LEE_NV_2	Mod_LEE_NV
male_US_EVAN_NV_2	Mod_EVAN_NV
male_BR_MALCOLM_NV_2	Mod_MALCOLM_NV
male_US_NATHAN_NV_2	Mod_NATHAN_NV
male_US_TOM_NV_2	Mod_TOM_NV

Tabla 7. Audios seleccionados de Nuance Vocalizer para generar modelos de locutor.

Se lanza la comparación de cada uno de ellos contra el resto de audios obtenidos con esta misma aplicación.

La población de referencia seleccionada consiste en los 35 modelos más competitivos de entre los 71 disponibles en voz de varón, en distintos canales, soporte digital y lengua inglesa. Esta población parece representar adecuadamente a los modelos de locutor en el caso de Lee y Malcolm, no así para el resto de modelos por lo que no ha sido posible seguir adelante con el cotejo en estos casos.

Los resultados obtenidos (expresados en valores de LR) se muestran en las tablas siguientes:

	Mod_LEE_NV
male_AU_LEE_NV_4	1.00E+10
male_AU_LEE_NV_1	1.00E+10
male_AU_LEE_NV_3	8.69E+09
male_BR_MALCOLM_NV_1	5.77E+07
male_BR_DANIEL_NV_3	1.95E+06
male_BR_MALCOLM_NV_2	9.49E+05
male_US_NATHAN_NV_2	8.18E+04
male_BR_MALCOLM_NV_3	6.27E+04
male_BR_DANIEL_NV_1	4.49E+03
male_BR_DANIEL_NV_2	3.60E+03
male_BR_MALCOLM_NV_4	2.90E+03
male_US_NATHAN_NV_3	5.78E+02
male_BR_DANIEL_NV_4	3.77E+02
male_US_NATHAN_NV_1	1.94E+01
male_US_EVAN_NV_3	1.81E+01
male_US_TOM_NV_2	1.30E+01
male_US_EVAN_NV_2	8.38E+00
male_US_TOM_NV_3	4.33E+00
male_US_NATHAN_NV_4	1.51E+00
male_US_EVAN_NV_4	9.00E-01
male_US_TOM_NV_1	7.05E-01
male_US_EVAN_NV_1	5.95E-01
male_US_TOM_NV_4	2.22E-01

Tabla 8. Resultados obtenidos de cotejo entre los archivos generados con Nuance Vocalizer y el modelo de locutor Mod_LEE_NV.

	Mod_MALCOLM_NV
male_BR_MALCOLM_NV_1	1.00E+10
male_BR_MALCOLM_NV_4	9.98E+09
male_BR_MALCOLM_NV_3	9.96E+09
male_AU_LEE_NV_2	1.44E+09
male_AU_LEE_NV_4	4.82E+08
male_AU_LEE_NV_1	2.29E+06
male_AU_LEE_NV_3	2.68E+05
male_US_EVAN_NV_2	1.08E+03
male_BR_DANIEL_NV_3	7.76E+02
male_US_NATHAN_NV_2	2.36E+02
male_BR_DANIEL_NV_2	4.63E+01
male_US_NATHAN_NV_3	2.57E+01
male_US_EVAN_NV_3	1.33E+01
male_BR_DANIEL_NV_1	8.90E+00
male_BR_DANIEL_NV_4	6.37E+00
male_US_TOM_NV_2	5.60E+00
male_US_EVAN_NV_1	4.20E+00
male_US_NATHAN_NV_1	3.44E+00
male_US_TOM_NV_3	1.46E+00
male_US_EVAN_NV_4	1.12E+00
male_US_NATHAN_NV_4	3.36E-01
male_US_TOM_NV_1	3.23E-01
male_US_TOM_NV_4	2.74E-01

Tabla 9. Resultados obtenidos de cotejo entre los archivos generados con Nuance Vocalizer y el modelo de locutor Mod_MALCOLM_NV.

Al igual que en el caso anterior la discriminación del sistema no es buena apareciendo puntuaciones por encima de uno para la mayoría de los audios que no pertenecen al mismo locutor.

No obstante, también se observa que los audios test pertenecientes al locutor elegido como modelo aparecen agrupados en la parte superior de la tabla con puntuaciones más altas que el resto. Incluso alguno de los audios de impostor presenta puntuaciones inferiores a uno.

5.3.3. Caso 3. Play HT.

Los audios seleccionados para generar modelos de locutor en este caso son:

ARCHIVO DE AUDIO	MODELO GENERADO
male_US_GUY_PHT	Mod_GUY_PHT
male_US_MARK_PHT_1	Mod_MARK_PHT
male_US_NOAH_PHT_1	Mod_NOAH_PHT

Tabla 10. Audios seleccionados de Play HT para generar modelos de locutor.

Se lanza la comparación de cada uno de ellos contra el resto de audios obtenidos con esta misma aplicación.

La población de referencia seleccionada consiste en los 35 modelos más competitivos de entre los 71 disponibles en voz de varón, en distintos canales, soporte digital y lengua inglesa.

En este caso no ha sido posible obtener resultados para ninguno de los modelos de locutor debido a que el sistema determina que la población de referencia no es suficientemente representativa.

6. Conclusiones y líneas de trabajo futuras.

En el planteamiento inicial de la fase experimental del trabajo se seleccionaron doce archivos con longitud adecuada para generar modelos de locutor, lo que habría reportado ese mismo número de comparaciones. Pero únicamente en cuatro de ellos el Sistema de Reconocimiento Automático ha determinado que las poblaciones de referencia de las que dispone son suficientemente representativas del modelo de locutor. Esto ha limitado considerablemente el número de comparaciones que ha sido posible realizar.

Este hecho, que a priori se podría considerar inconveniente, también nos aporta información útil acerca de cómo responde nuestro sistema frente a este tipo de habla: **las poblaciones de referencia de voz natural no son, en la mayoría de los casos, representativas de los modelos de locutor de voz artificial.**

En todos los casos en los que ha sido posible realizar las comparaciones e independientemente de la aplicación, los resultados presentan la misma tendencia: los valores de LR son en su mayoría superiores a uno lo que, sin hacer una interpretación detallada, significaría que todos audios pertenecen al mismo locutor y por tanto **la respuesta del sistema no parece buena.**

Detrás de ello estaría la posible influencia de los procesos que las aplicaciones utilizan para generar las grabaciones, que podrían aportar a la estructura de la señal acústica algún elemento que el sistema reconoce como una característica común a todos los hablantes. Esto podría interpretarse como una **prevalencia de los procesos utilizados por la aplicación en la generación de la voz frente a las características intrínsecas de la misma.**

Además, es preciso tener en cuenta que el desajuste de las poblaciones de referencia respecto a nuestros modelos de sospechoso repercute en que todas las puntuaciones sean más altas de lo que deberían.

No obstante, **sí se observa cierto grado de discriminación** al considerar la magnitud de las puntuaciones. En general, los audios test del locutor que coincide con el modelo aparecen agrupados en la parte alta de la tabla, con los mayores valores de

LR. Así, por ejemplo, para el locutor de inglés australiano de la aplicación SPIK-AI se obtienen valores del orden de 10^{10}, y para el resto de locutores los valores están en el orden de 10^{8} a 10^{-2}. Lo mismo ocurre en el caso de los locutores Malcom y Lee de la aplicación NUANCE VOCALIZER. Estos resultados **no son incoherentes desde el punto de vista de la capacidad discriminativa del sistema**.

Todos estos condicionantes hacen no recomendable la utilización de un Sistema de Reconocimiento Automático de Locutores como única aproximación de estudio con grabaciones de este tipo.

Los ensayos realizados no serían más que una primera aproximación al análisis de voces sintéticas en el ámbito forense, sin más pretensión que observar el comportamiento del SRAL ante ellas. Más allá del uso de los SRAL, las distintas aproximaciones de estudio ofrecen la posibilidad de realizar análisis más detallados que nos permitan detectar características propias de grabaciones de voz artificial.

Yendo un paso más allá, lo verdaderamente interesante sería repetir en el futuro este tipo de estudio utilizando aplicaciones de clonación de voz. Además de las aplicaciones comercializadas por las grandes empresas, existen en el mercado varias opciones que permiten a los usuarios con un nivel medio de conocimiento de ciertos lenguajes de programación crear su propia herramienta de clonación, lo que nos ofrece la posibilidad de comparar la voz natural original con la voz clonada obtenida a partir de ella.

7. Bibliografía.

[1] Cruzcampo | Así se hizo #ConMuchoAcento, YouTube, 2021. https://www.youtube.com/watch?v=BQLTRMYHwvE (consultado el 6 de septiembre de 2021).

[2] C. D. Romero, La identificación de locutores en el ámbito forense, tesis, Universidad Complutense de Madrid, 2001.

[3] A. Bonavida, Boletín de la Asociación Española de Logopedia, Foniatría y Audiología, (s.f.).

[4] L. M. Carrillo, Influencia del plano expresivo en las estructuras acústicas de resonancia del habla, Trabajo Fin de Máster, Instituto Universitario de Investigación en Ciencias Policiales, Universidad de Alcalá de Henares, 2020.

[5] M. R. L. Avilés, Utilidad del análisis sonográfico de banda estrecha en la caracterización de rasgos acústicos del habla, Trabajo Fin de Máster, Instituto Universitario de Investigación en Ciencias Policiales, Universidad de Alcalá de Henares, 2016.

[6] A. D. G. Sánchez, El reto de los nuevos formatos de audio multimedia en Acústica Forense, Trabajo Fin de Máster, Departamento de Ciencia y Técnica Policial, Escuela Nacional de Policía, 2019.

[7] Manual de usuario Batvox 3.0 Pro, Agnitio Voice Biometrics, 2009.

[8] Síntesis de habla, Wikipedia. (2021). https://es.wikipedia.org/wiki/S%C3%ADntesis_de_habla (consultado el 6 de septiembre de 2021).

[9] El falso Barack Obama creado con inteligencia artificial capaz de hablar como si fuera el original, BBC News, 2017. https://www.bbc.com/mundo/media-40632577 (consultado el 7 de septiembre de 2021).

[10] La inteligencia artificial resucita la voz de Franco, El País. (2020). https://elpais.com/tecnologia/2020-06-02/la-inteligencia-artificial-resucita-la-voz-de-franco.html (consultado el 7 de septiembre de 2021).

[11] Así es como Franco ha 'resucitado' para cantar la Macarena, El Español. (2020). https://www.elespanol.com/omicrono/tecnologia/20200616/franco-resucitado-cantar-macarena/497951283_0.html (consultado el 7 de septiembre de 2021).

[12] C. Villoria, Reconocimiento y síntesis de voz, Recursostic.educacion.es. (2009). http://recursostic.educacion.es/observatorio/web/es/software/software-general/689-reconocimiento-y-sintesis-de-voz (consultado el 4 de septiembre de 2021).

[13] I. Iriondo, J. Martí, J. Oliver, R. Guaus, H. Moure, Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla, Dept. De Comunicacions i Teoria Del Senyal. Enginyeria La Salle. Universitat Ramon Llull. (1999).

[14] A.B. Cávez, Tecnologías del habla: Conversión de texto a voz, Departament De Teoria Del Senyal i Comunicacions, UPC. (1997).

[15] Síntesis articulatoria, Wikipedia. (2021). https://tvd.wiki/wiki/Articulatory_synthesis (consultado el 7 de septiembre de 2021).

[16] M.Á.R. Crespo, Introducción a la conversión texto-voz, Philologia Hispalensis. (1997). 177–192.

[17] J.D.V. García, Redes NEURONALES Desde cero (i) - Introducción, IArtificial.net. (2020). https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion (consultado el 6 de septiembre de 2021).

[18] F.S. Caparrini, Redes Neuronales: Una visión superficial, Redes Neuronales: Una Visión Superficial - Fernando Sancho Caparrini. (2019). http://www.cs.us.es/~fsancho/?e=72 (consultado el 6 de septiembre de 2021).

[19] E.R. Álvarez, ¿Por Qué el deep LEARNING ha superado al machine learning?, Thinking for Innovation. (2021). https://www.iebschool.com/blog/machine-learning-deep-learning-big-data/ (consultado el 6 de septiembre de 2021).

[20] P.P. Torralba, ¿Qué es machine learning (aprendizaje automático)? Thinking for Innovation. (2021). https://www.iebschool.com/blog/que-machine-learning-big-data/ (consultado el 6 de septiembre de 2021).

[21] G. Julián, Las redes Neuronales: Qué Son y por qué están volviendo, Xataka. (2016). https://www.xataka.com/robotica-e-ia/las-redes-neuronales-que-son-y-por-que-estan-volviendo (consultado el 6 de septiembre de 2021).

[22] F. Izaurieta, C. Saavedra, Redes Neuronales Artificiales, Departamento De Física, Universidad De Concepción. (2000).

[23] A. González, ¿Qué es machine learning?, Cleverdata. (2014). https://cleverdata.io/que-es-machine-learning-big-data/ (consultado el 6 de septiembre de 2021).

[24] Redes NEURONALES Profundas: Qué Son y cómo funcionan, Psicología y Mente. (2021). https://psicologiaymente.com/cultura/redes-neuronales-profundas (consultado el 6 de septiembre de 2021).

[25] J.D.V. García, Redes NEURONALES Desde cero (II): Algo de matemáticas, IArtificial.net. (2020). https://www.iartificial.net/redes-neuronales-desde-cero-ii-algo-de-matematicas (consultado el 6 de septiembre de 2021).

[26] T. Rodríguez, Machine learning y deep learning: Cómo entender las claves del presente y futuro de la inteligencia artificial, Xataka. (2020). https://www.xataka.com/robotica-e-ia/machine-learning-y-deep-learning-como-entender-las-claves-del-presente-y-futuro-de-la-inteligencia-artificial (consultado el 6 de septiembre de 2021).

[27] Iberdrola Corporativa, Descubre Los PRINCIPALES beneficios del machine learning, Iberdrola. (n.d.). https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico (consultado el 6 de septiembre de 2021).

[28] Idiomas y voces, Documentos De IBM Cloud. (2020). https://cloud.ibm.com/docs/text-to-speech?topic=text-to-speech-voices&locale=es (consultado el 6 de septiembre de 2021).

[29 M. Rodríguez, Google presenta Tacotron 2, un sistema de IA con una voz prácticamente humana, 1million bot. (2017). https://1millionbot.com/tacotron-2-google-ia-con-voz-humana/ (consultado el 8 de septiembre de 2021).

[30] R. Arrabales, Deep learning: Qué es y por qué va a ser una tecnología clave en el futuro de la inteligencia artificial, Xataka. (2016). https://www.xataka.com/robotica-e-ia/deep-learning-que-es-y-por-que-va-a-ser-una-tecnologia-clave-en-el-futuro-de-la-inteligencia-artificial (consultado el 6 de septiembre de 2021).

[31] Spik.AI. https://spik.ai/

[32] Nuance Vocalizer. https://www.nuance.com/es-es/omni-channel-customer-engagement/voice-and-ivr/text-to-speech.html

[33] Play HT. https://play.ht/

INSTITUTO UNIVERSITARIO DE INVESTIGACIÓN EN CIENCIAS POLICIALES

Facultad de Derecho. Universidad de Alcalá
Libreros, 27, planta baja. 28801 Alcalá de Henares (Madrid)
https://iuicp.uah.es