FRECUENCIAS DE LAS SECUENCIAS SILÁBICAS PARA LA VERIFICACIÓN DE HABLANTES EN TRANSCRIPCIONES DE AUDIOS

Johny Alexander Cárdenas Rodríguez

Facultad Seminario Andrés Bello, Instituto Caro y Cuervo
Maestría en Lingüística

Dr. Sergio Gonzalo Jiménez Vargas

Bogotá D.C.

Dedicatoria

El arduo esfuerzo requerido para el desarrollo de este trabajo fue logrado gracias al apoyo de mi familia en los momentos más difíciles. Por ello, este trabajo va para mi madre quien me ayudó a lo largo de este nuevo camino.

Agradecimientos

A los profesores del Instituto Caro y Cuervo quienes transmitieron el conocimiento de un área tan compleja como es la Lingüística. En mención especial al profesor Sergio Jiménez quien fue el faro que me ayudó al desarrollo adecuado del presente trabajo de grado.

	AUTORIZACIÓN DEL AUTOR PARA CONSULTA Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE GRADO	Código: FOR-F-2 Versión: 1.0 Página 1 de 1 Fecha: 17/03/2022

BIBLIOTECA JOSÉ MANUEL RIVAS SACCONI

INFORMACION DEL TRABAJO DE GRADO

1. **Trabajo de grado requisito para optar al título de: Maestro en Lingüística**

2. **Título del trabajo de grado: Frecuencias de las secuencias silábicas para la verificación de hablantes en transcripciones de audios**

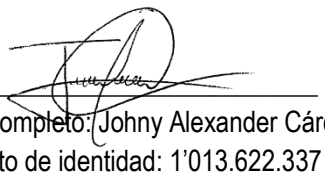
3. **Autoriza la consulta y publicación electrónica del trabajo de grado:**

Sí autorizo [x] **, No autorizo** [ ] **a la biblioteca José Manuel Rivas Sacconi del Instituto Caro y Cuervo para que con fines académicos:**

- Ponga el contenido de este trabajo a disposición de los usuarios en la biblioteca digital Palabra, así como en redes de información del país y del exterior, con las cuales tenga convenio la Facultad Seminario Andrés Bello y el Instituto Caro y Cuervo.
- Permita la consulta a los usuarios interesados en el contenido de este trabajo, para usos de finalidad académica, ya sea formato impreso, CD-ROM o digital desde Internet.
- Socialice la producción intelectual de los egresados de las Maestrías del Instituto Caro y Cuervo con la comunidad académica en general.
- Todos los usos, que tengan finalidad académica; de manera especial la divulgación a través de redes de información académica.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, ***"Los derechos morales sobre el trabajo son propiedad de los autores"***, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. Atendiendo lo anterior, siempre que se consulte la obra, mediante cita bibliográfica se debe dar crédito al trabajo y a su autor.

4. **Identificación del autor**

Firma: 

Nombre completo: Johny Alexander Cárdenas Rodríguez

Documento de identidad: 1'013.622.337 de Bogotá D.C.

DESCRIPCIÓN TRABAJO DE GRADO

AUTOR

Apellidos	Nombres
Cárdenas Rodríguez	Johny Alexander

DIRECTOR (ES)

Apellidos	Nombres
Jiménez Vargas	Sergio Gonzalo

TRABAJO PARA OPTAR POR EL TÍTULO DE: Maestro en Lingüística

TÍTULO DEL TRABAJO DE GRADO: Frecuencias de las secuencias silábicas para la verificación de hablantes en transcripciones de audios

NOMBRE DEL PROGRAMA ACADÉMICO: Maestría en Lingüística

CIUDAD: _Bogotá_ AÑO DE PRESENTACIÓN DEL TRABAJO: 2024

NÚMERO DE PÁGINAS:

TIPO DE ILUSTRACIONES: Ilustraciones __ Mapas ___ Retratos ___ Tablas, gráficos y diagramas _X__ Planos ___ Láminas ___ Fotografías ___

MATERIAL ANEXO (Vídeo, audio, multimedia):

Duración del audiovisual: _0_ Minutos.

Otro. ¿Cuál? _____

Sistema: Americano NTSC _____ Europeo PAL _____ SECAM _____

PREMIO O DISTINCIÓN (En caso de ser Laureadas o tener una mención especial):
Laureada

DESCRIPTORES O PALABRAS CLAVES: Son los términos que definen los temas que identifican el contenido. *(En caso de duda para designar estos descriptores, se recomienda consultar a la dirección de biblioteca en el correo electrónico biblioteca@caroycuervo.gov.co):*

ESPAÑOL **INGLÉS**

frecuencia silábica, silabario mental,	syllable frequency, mental syllabary,
transiciones de sílabas, verificación forense	syllable transitions, forensic speaker
de hablantes, comparación de hablantes,	verification, speaker comparison,
transcripciones	transcriptions

RESUMEN DEL CONTENIDO Español (máximo 250 palabras):

La comparación forense de hablantes utiliza varios métodos de análisis, algunos basados en atributos fisiológicos de la voz. Sin embargo, la calidad del audio puede afectar la precisión de las mediciones segmentales y suprasegmentales debido a factores adversos.

Este estudio propone un enfoque centrándose en las secuencias silábicas de las transcripciones de las grabaciones. El método se basa en el Efecto de Frecuencia de Sílabas, que demuestra la existencia del Silabario Mental con eficiencia diferencial por frecuencia silábica. Al representar el habla espontánea por las transiciones de frecuencias de las sílabas, se codifica el acceso eficiente al Silabario Mental de cada hablante, determinado por factores individuales y colectivos.

Para verificar esta hipótesis, se construyó un sistema de verificación de hablantes basado en esta representación, usando regresión logística, frecuencias silábicas de Twitter (ahora X) y muestras de habla de los conjuntos de datos AusEng500 y ForVoice120. Los resultados superaron ampliamente los métodos basados en transcripciones de audios y fueron comparables con métodos de audio y aprendizaje profundo. Además, a diferencia de los modelos no interpretables del estado del arte, se presentó una visualización de "firmas de transiciones espectrales silábicas" que puede diferenciar hablantes a simple vista.

Concluimos que la representación del habla espontánea mediante transiciones espectrales silábicas es útil para construir sistemas de verificación forense de hablantes de alto rendimiento, y que las "firmas de transiciones espectrales silábicas" abren una nueva perspectiva para la admisibilidad judicial de pruebas lingüísticas.

RESUMEN DEL CONTENIDO Inglés (máximo 250 palabras):

Forensic speaker comparison uses several methods of analysis, some based on physiological attributes of the voice. However, audio quality can affect the accuracy of segmental and suprasegmental measurements due to adverse factors.

This study proposes an approach focusing on the syllabic sequences of the transcriptions of the recordings. The method is based on the Syllable Frequency Effect, which demonstrates the existence of the Mental Syllabary with differential efficiency by syllable frequency. By representing spontaneous speech by syllable frequency transitions, the efficient access to the Mental Syllabary of each speaker, determined by individual and collective factors, is encoded.

To verify this hypothesis, a speaker verification system based on this representation was built using logistic regression, Twitter (now X) syllable frequencies and speech samples from the

AusEng500 and ForVoice120 datasets. The results far outperformed methods based on audio transcripts and were comparable with audio and deep learning methods. Furthermore, unlike state-of-the-art non-interpretable models, a visualization of “signatures of syllabic spectral transitions” was presented that can differentiate speakers at a glance.

We conclude that the representation of spontaneous speech by syllabic spectral transitions is useful for building high-performance forensic speaker verification systems, and that “signatures of syllabic spectral transitions” open a new perspective for the judicial admissibility of linguistic evidence.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Resumen

La comparación forense de hablantes emplea diversos métodos de análisis y algunos están relacionados con atributos fisiológicos de la voz. Sin embargo, la calidad del audio afecta la precisión de las mediciones de los elementos segmentales y suprasegmentales debido a factores adversos por la naturaleza de la tarea.

En contextos forenses, la calidad de los archivos de audio utilizados en los laboratorios policiales es a menudo regular o mala. Estos audios suelen presentar altos niveles de ruido de fondo, tiempos de reverberación elevados, y compresión de la señal. Estos y otros factores humanos, como la impostación y la alteración de la fonación debido al uso de elementos como el pinzamiento de la nariz modifican la voz e influyen directamente sobre las características acústicas. Es por ello que, el presente estudio plantea un enfoque independiente de los parámetros acústicos para la verificación de hablantes, centrándose en el elemento idiolectal de las secuencias silábicas en las transcripciones de las grabaciones.

El método está motivado por el Efecto de Frecuencia de Sílabas (Cholin, Dell & Levelt, 2011), que demostró la existencia del Silabario Mental con eficiencia diferencial por frecuencia silábica. Así, al representar el habla espontánea por sus transiciones de rangos de frecuencias de las sílabas, se codifica el acceso eficiente al Silabario Mental de cada hablante. Si este acceso eficiente está determinado por factores individuales como la adquisición de la lengua y las experiencias sensorio-motoras, y por otra parte, por factores colectivos como la educación y las relaciones sociales, entonces la representación propuesta tendría carácter individualizante.

Para verificar esta hipótesis, construimos un sistema de verificación de hablantes basado en dicha representación usando: regresión logística, frecuencias silábicas obtenidas de

Twitter (ahora X) y muestras de habla de los conjuntos de datos para verificación forense AusEng500 (en inglés) y ForVoice120+ (en húngaro). Los resultados superaron ampliamente los métodos en el estado del arte basados en transcripciones de audios, y fueron comparables con métodos basados en audio y aprendizaje neuronal profundo (*Deep Learning*). Adicionalmente, a diferencia del estado del arte donde los modelos no son interpretables, usando los coeficientes de la regresión logística, presentamos una visualización de "firmas de transiciones espectrales silábicas" la cual tiene el potencial de diferenciar hablantes a simple vista.

Concluimos que la representación del habla espontánea usando transiciones espectrales silábicas es útil para construir sistemas de verificación forense de hablantes de alto rendimiento, y que las "firmas de transiciones espectrales silábicas" abren una nueva perspectiva hacia la admisibilidad judicial de pruebas de laboratorio lingüísticas.

Palabras clave: frecuencia silábica, silabario mental, transiciones de sílabas, verificación forense de hablantes, comparación de hablantes, transcripciones.

1. Introducción

El proceso de verificación forense de hablantes es una tarea de suma importancia para el sistema judicial debido a que su función es la de responder a la pregunta: ¿qué tan probable es que las muestras de habla indubitada (muestra de origen conocido) y dubitada (muestra de origen desconocido) provengan de una misma persona? (San Segundo et. al., 2019). Para abordar esta cuestión, los científicos forenses han desarrollado metodologías que evalúan aspectos acústicos mediante el uso de herramientas automáticas o semiautomáticas. Estas herramientas proporcionan resultados de probabilidad que respaldan ya sea la hipótesis de la

Fiscalía (H_0) de que ambas muestras proceden del mismo hablante o la hipótesis de la contraparte (H_1) de que provienen de hablantes distintos.

Los métodos de comparación de hablantes realizan mediciones de distintos parámetros acústicos, que van desde la frecuencia fundamental F_0 hasta aspectos prosódicos como la entonación (San Segundo et. al., 2018). Estos métodos proporcionan resultados cuantitativos basados en una razón de verosimilitud (*Likelihood Ratio*). Sin embargo, el análisis cuantitativo de los parámetros es dependiente de la calidad de la señal, la cual a menudo se ve afectada por diversos factores externos como altos niveles de ruido ambiental, de reverberación o también factores propios del hablante al realizar cambios intencionados sobre la voz (San Segundo et. al., 2019).

En los métodos semiautomáticos, se complementa el análisis acústico automático con la evaluación de las características lingüísticas observando marcas lingüísticas distintivas que el hablante presenta en su oratoria (Romero, 2001). La herramienta que permite el análisis de estos elementos es la elaboración de un "corpus lingüístico" de la persona. Esta elaboración consiste en realizar una transcripción buscando rescatar la mayor información lingüística. La extracción y análisis de las características lingüísticas se logra gracias a la experticia y habilidad auditiva entrenada de los peritos forenses. Por lo tanto, es posible, para un experto, discriminar las palabras incluso en ambientes con un alto nivel de ruido o con características adversas del audio. La habilidad de centrar la atención a una fuente se le conoce como *cocktail party effect* (Arons, 1992), la cual consiste en la lograr aislar de manera perceptiva las emisiones de un solo hablante a pesar del ruido o de las conversaciones circundantes.

En consecuencia, los peritos pueden lograr extraer una gran cantidad de información, lo que resulta en la obtención de transcripciones fidedignas que tienen el potencial

discriminatorio y pueden ser utilizadas para el estudio comparativo. La transcripción se lleva a cabo para obtener un corpus que permita recolectar información sobre las emisiones comunicativas (Erickson, 2017), siendo una herramienta de ubicación temporal para el audio y de extracción de características lingüísticas relevantes.

El corpus generado a partir de las transcripciones posee un gran potencial de estudio, ya que se pueden aplicar herramientas tradicionalmente utilizadas en la atribución de autoría en textos escritos para lograr validar la identidad de una persona. Es por esta razón, que el presente estudio tiene como objetivo mostrar que cada individuo en su discurso tiene patrones individualizantes[1] influenciados por su pensamiento, modelando así, su lenguaje que lo distingue del resto de la población. Es por ello que, en principio, cada hablante (o escritor) posee un idiolecto propio, que opera de manera inconsciente, y que se encuentra condicionado por sus hábitos personales, interacciones sociales, formación educativa y experiencias que ha tenido a lo largo de su vida (Hernández, 1980).

Si el idiolecto opera de manera inconsciente, entonces se implica que pueden existir factores cognitivos que lo afectan. En el habla espontánea, el discurso es planeado y producido de manera eficiente y en tiempo real por el hablante involucrando aspectos lingüísticos y cognitivos de la comunicación (Simpson, 2013). Uno de los aspectos cognitivos que afecta la planeación y articulación de las palabras es el Efecto de Frecuencia de Sílabas (*Syllable-Frequency Effect*), que indica que los hablantes producen sílabas con diferente eficiencia de acuerdo a sus frecuencias (Cholin, Dell y Levelt, 2011). Este efecto es la primera evidencia de la existencia del Silabario Mental (*Mental Syllaraby*) el cual es accedido previamente a la codificación fonética (Levelt & Wheeldon, 1994). Adicionalmente, esta

[1] En el artículo se expone los resultados únicamente de la secuencia silábica y su frecuencia en la cadena hablada.

existencia se ha corroborado con evidencia fisiológica usando Imágenes de Resonancia Magnética Funcional (Brendel et al., 2011). Es así, que la eficiencia con la que los hablantes producen las sílabas está correlacionada con las frecuencias en un corpus representativo. Aunque aún no se conoce cómo se organiza y almacena el Silabario Mental en la cognición, es plausible hacer una analogía con el Lexicón Mental, el cual no se almacena como una simple lista ordenada sino en una estructura compleja interconectada (Schiller, 2021). Durante el habla espontánea, la secuencia de sílabas se produce de manera natural haciendo un uso intensivo y eficiente del Silabario Mental y de su estructura de almacenamiento mental. Si la estructura de almacenamiento del Silabario Mental es diferente en cada individuo, sería posible observar diferentes patrones en las secuencias de sílabas y sus frecuencias producidas durante el habla espontánea. Es precisamente este factor psicolingüístico el que se explora en este estudio con el objetivo de diferenciar e identificar hablantes basándose en secuencias de sílabas extraídas de muestras de habla y en frecuencias silábicas obtenidas de corpus representativos de los hablantes.

Basados en esta motivación, presentamos una nueva representación vectorial para el habla la cual denominamos “transiciones espectrales silábicas”. La idea general es representar las sílabas por sus frecuencias y luego construir un vector usando conteos de las transiciones de dichas frecuencias. La primera pregunta que buscamos responder en este estudio es si esta representación vectorial inspirada en el Efecto de Frecuencia de las Sílabas y el Silabario Mental es útil para individualizar a los hablantes. El método para dar respuesta a esta pregunta consiste en construir un sistema de identificación de hablantes en ambiente forense usando esa representación vectorial y evaluar su rendimiento (Beke, 2021 y Morrison G.S. et. al., 2021). Adicionalmente, este rendimiento será comparado con el de otros sistemas del estado del arte

basados en otros tipos de datos (i.e. audios) y de representaciones del habla transcrita (i.e. semánticas). Para una posible respuesta positiva a la pregunta, el sistema de identificación propuesto deberá tener un rendimiento comparable o superior.

Otro problema de la verificación forense de hablantes basada en la voz es la admisibilidad de sus juicios en estrados judiciales (Morrison, 2014). Desde los años 1960s estos juicios han estado evolucionando hasta la actualidad; desde el análisis manual y minucioso de casos individuales, hasta la validación empírica usando grandes cantidades de datos (Morrison et al., 2021). Los modelos automáticos actuales para esto último, se basan en modelos de redes neuronales con grandes cantidades de parámetros, lo que en la práctica impide su interpretabilidad y por ende la producción de pruebas admisibles (Garret y Rudin, 2023). Las secuencias de sílabas son una representación mucho más compacta hasta en tres órdenes de magnitud que las secuencias de audio, haciendo que sea mucho más factible construir modelos de verificación de hablantes interpretables. La segunda pregunta a abordar en este estudio es investigar si a partir del uso de las Transiciones Espectrales Silábicas se puede obtener pruebas de identidad de hablantes interpretables. El método para determinar esto será utilizar para la verificación forense un modelo de clasificación interpretable como la regresión logística y construir una visualización de sus parámetros principales. Para una posible respuesta positiva a la pregunta, esta visualización deberá permitir distinguir claramente a los hablantes por simple inspección visual.

2. Marco teórico

2.1. Fundamentos de la variación lingüística

El análisis de atribución forense de autoría se basa en la premisa de que cada individuo posee su propio modo de utilizar el lenguaje, conocido como idiolecto, el cual lo distingue de

los demás. Por consiguiente, la variación lingüística está intrínsecamente ligada a una serie de factores que dan lugar a diferentes formas de expresión dentro de los sistemas lingüísticos. Las corrientes estructuralista, funcionalista y generativista han formulado postulados sobre esta variación lingüística entre los hablantes.

Saussure (1978), en su influyente obra “Curso de lingüística general”, establece los principios del estructuralismo lingüístico. En sus tratados argumenta que el lenguaje se puede representar como un sistema de elementos lingüísticos estructurados los cuales están destinados a ocupar determinadas posiciones específicas en el habla, formando así los signos que determinan las relaciones de los elementos y excluyen otras posibles combinaciones para la comunicación. Según esta perspectiva, las combinaciones lingüísticas son sistemáticas en la estructura de la lengua y se generan como una consecuencia contextual del significado expresado a través del lenguaje. Así, el análisis lingüístico determina las características idiosincráticas posibles que son limitadas por las reglas lingüísticas. No hay razones particulares para pensar que el lenguaje está limitado de esta manera pero sí muestra que el habla está condicionada por estructuras coherentes.

Desde la perspectiva del funcionalismo, se reconocen múltiples aspectos de la variación lingüística, que incluyen la diafasía, la diastratía y la diatopía. Estos niveles se ocupan de las diferencias en la expresión o las modalidades expresivas en la dimensión estilística del lenguaje. Por tanto, el estilo de habla de una persona está estrechamente relacionado con la situación comunicativa y el uso de los rasgos connotativos o marcas pragmáticas que generan los usos del lenguaje (Casa Gómez, 2008).

Ahora bien, el generativismo, siguiendo la línea de pensamiento de Noam Chomsky (1965), establece la relación entre el fenómeno de la variación lingüística y la competencia

gramatical del hablante. Esta competencia se manifiesta en la aplicación de reglas y principios específicos que forman parte de la gramática interna de cada individuo, lo que permite generar un amplio conjunto de oraciones que son gramaticalmente correctas. A partir de esta perspectiva, el lenguaje se concibe como el resultado de la influencia de factores internos que rigen la selección de las palabras y las reglas sintagmáticas frente a otras.

La vertiente variacionista se enfoca en concebir el concepto de la lengua como un resultado de la interacción social (Labov, 1972), indicando que la variación lingüística es una propiedad inherente que no solo está regida por factores internos (descritos en los enfoques estructuralistas y generativistas) sino también por factores externos (expuesto por la corriente funcionalista). Los factores externos ejercen influencia en la variación lingüística dentro de las comunidades de hablantes de una lengua en particular. El análisis cualitativo y cuantitativo de la variación revela que los miembros de comunidades distintas, ya sea en términos geográficos o sociales, no comparten los mismos rasgos lingüísticos. A nivel individual, esta variación se puede contextualizar de acuerdo con factores como la ideología, la fisiología y los hábitos lingüísticos de cada hablante.

2.2. El lenguaje a como expresión individual

A nivel psicolingüístico, la producción de palabras se define como un proceso de múltiples etapas que va desde la representación conceptual (ligada a la familiaridad del concepto) hasta las entradas léxicas donde la persona selecciona aquella palabra o *lemma* (entrada léxica abstracta que contiene información sintáctica) adecuada, culminando con la articulación fonética (Wilson et al 2009). Estas etapas suelen desarrollarse sin ningún esfuerzo y de manera inconsciente.

Ahora bien, la teoría de la variación lingüística (Labov, 1972), proporciona la base metodológica y conceptual esencial para los estudios de comparación lingüística de individuos. Según esta teoría, cada hablante nativo de una lengua posee su propia versión individual que es distintiva. En otras palabras, los hablantes son conscientes de la independencia léxica y tienen un control total sobre las elecciones que hacen en el proceso de producción lingüística, lo que da lugar a su propio idiolecto. A su vez, este idiolecto se manifiesta a través de selecciones singulares en los distintos niveles lingüísticos al producir un enunciado, incluyendo los aspectos fonéticos, léxicos, sintácticos y semánticos. Como resultado, las elecciones son constantes y pueden detectarse tanto en sus expresiones orales como escritas. Esta premisa se utiliza en el ámbito forense para abordar las comparaciones, destacando que, en teoría, cada persona posee una forma única de expresarse debido a la estrecha relación entre el lenguaje y el pensamiento.

Los estudios de la relación entre el lenguaje y el pensamiento han sido desarrollados por la psicología experimental al ver los procesos de adquisición del lenguaje. Piaget sostiene que el desarrollo de la cognición está intrínsecamente relacionado con la evolución de esquemas sensorio-motores encargados de organizar las experiencias, de manera que los pensamientos tienen su raíz en la acción individual y que es debido a los aprendizajes propios de cada individuo. Ahora bien, la adquisición del lenguaje surge a partir de las experiencias sensoriales y, según Piaget, esto marca el inicio de la esquematización representativa individual de las personas al estar condicionado por la acción aunque esta es una parte para la constitución de las operaciones mentales.

Hernández (1980) expone el punto de vista de Vygotsky donde indica que la actividad mental es el resultado de las circunstancias sociales, y por tanto, está condicionada por la

comunicación que el individuo desarrolla con sus redes sociales. En consecuencia, las experiencias son el producto de la intercomunicación. Vygotsky sostiene que el lenguaje tiene dos funciones: la comunicación externa con los demás y la manipulación de los pensamientos internos. Ahora bien, Bruner ha llegado a un acuerdo con ambas posturas formulando que el lenguaje está organizado de modo jerárquico según la experiencia y según las relaciones sociales.

Entonces, de manera teórica, el pensamiento logra encontrar su forma a través del lenguaje y, a su vez, predispone al individuo al uso de su repertorio lingüístico basado en las experiencias e interacciones comunicativas para luego reproducirlas en sus enunciados. Por tanto, los rasgos propios de su lenguaje son susceptibles de análisis y permiten visibilizar la selección léxica junto con la preferencia de la posición en la cadena hablada.

2.3. Efecto de Frecuencias de Palabras

Además de la capacidad de cada persona de seleccionar las palabras en su oratoria, es requerido analizar la cantidad de veces que utiliza las palabras según su familiaridad y preferencia. Según Brysbaert (2018), la frecuencia de palabras enmarca los aspectos paralingüísticos cognitivos de las personas relacionando su capacidad argumentativa con la producción que realiza al momento de expresarla. Las investigaciones de la frecuencia de palabras, desde un enfoque cognitivo, han encontrado que no todas las palabras son procesadas en el cerebro con la misma eficiencia (Brysbaert et. al., 2011). Las palabras de una mayor frecuencia son conocidas por más personas y procesadas de manera más rápida que las de frecuencias más bajas.

En el contexto de este proyecto, se hace uso indirecto del Efecto de la Frecuencia de las Palabras (EFP) aplicado en las unidades silábicas para la caracterización de los hablantes

(Brysbaert, 2018). Para que el EFP pueda ser perceptible, es necesario que el corpus extraído de un hablante, sea relevante para que pueda existir una correlación con la riqueza léxica individual. Así mismo, para cuantificar las palabras extraídas, es necesario contar con información del área de influencia lingüística o zona geográfica circundante, permitiendo visualizar aquellos aspectos que son propios del hablante de estudio y que lo diferencian del resto de la comunidad.

Es así, como el EFP se conceptualiza como la producción de las palabras de un individuo correlacionadas con su ocurrencia en un corpus base que registra la cantidad de veces que la palabra aparece en mayor o menor medida en la zona de influencia. Esta relación está vinculada a la familiaridad de los conceptos de las personas y, por ende, está asociada con la activación cognitiva del *lemma*. Estudios en psicología experimental han demostrado que esta frecuencia es una variable relevante para las decisiones léxicas de las personas, influenciada con la familiaridad o el conocimiento experimental según su entorno, lo cual se refleja en la conformación de las unidades léxicas (Brysbaert, 2018).

Debido a que no es posible acceder a corpus de gran tamaño en modalidad oral para el húngaro y el inglés, se utiliza como herramienta corpora proveniente de Twitter (ahora X) y otras fuentes en línea. Los *trinos* son apoyados con las frecuencias extraídas de periódicos y blogs consignados por Gimenes et. al. (2015) en el estudio de las decisiones léxicas, otorgando la información suficiente para las tablas de frecuencias en inglés y húngaro.

2.4. Efecto de Frecuencia de Sílabas

Los elementos lingüísticos se almacenan en la mente de manera estructurada y se configuran en función de su frecuencia de uso y familiaridad. Aquellos que son más cortos o más familiares para el hablante o escritor tienden a tener una mayor preferencia en su

comunicación. Por consiguiente, se puede observar una generalización de patrones morfológicos que da lugar a una red asociativa donde se guardan las formas y combinaciones (Bybee, 2001). Esta realidad lleva a las personas a activar estructuras lingüísticas basadas en su conocimiento del uso de las sílabas para construir palabras que se ajusten al contexto. En el momento de formular enunciados, los morfemas se agrupan según la organización silábica de cada unidad léxica.

En el habla espontánea, los sonidos individuales se juntan para formar unidades silábicas que son acordes a las reglas lingüísticas de la lengua para su comprensión. Tan pronto como un hablante ha seleccionado una unidad fraseológica en su léxico mental, se inicia la codificación fonológica. Esta codificación consiste en un conjunto ordenado de segmentos que conforman el inventario silábico mental el cual gestiona la posterior codificación fonética, la cual se encarga de modular a los órganos fonoarticuladores para la producción oral.

Modelos de la producción del habla propuestos por Levelt et. al. (1999) han instaurado una base teórica del acceso léxico al momento de emitir las cadenas habladas, y en ellas las sílabas juegan un papel importante como la interface entre la generación abstracta fonología y su realización fonética (Cholin et. el., 2005). Es entonces, cuando las sílabas no son un constructo desarrollado en cada momento del habla sino que son, hipotéticamente, precompilados gesticulares para el movimiento de los articuladores. En los estudios de Cholin, se logra evidenciar que ese inventario silábico existe en términos de frecuencia de sílabas. La facilidad de producción y ejecución silábica depende de la familiaridad con los conceptos e incluso los hablantes suelen reutilizar un pequeño conjunto de sílabas una y otra vez (Schiller

et al., 1996), por lo que la frecuencia de las sílabas supone una característica marcada de los hablantes.

2.5. Transiciones léxicas

Según Chomsky (1965), la capacidad lingüística tiene un carácter innato en el proceso de adquisición del lenguaje el cual es interactivo con el entorno lingüístico. Esta interacción entre lo conocido (conocimiento previo) y lo adquirido resulta en la formación gramatical propia que se rige a partir de la relación semántica y sintáctica, buscando un orden correcto establecido a partir de las enseñanzas y experiencias. La relevancia de esta teoría es crucial ya que aborda la adquisición de la capacidad sintáctica que permite demostrar el carácter propio de las secuencias de palabras y, de manera derivada, las secuencias silábicas, que son el objeto del presente estudio.

Siguiendo la teoría de Chomsky, una persona a partir que empieza a adquirir el lenguaje, puede derivar las reglas gramaticales de manera adecuada, creando oraciones estructuradas que tienen sentido, se demuestra que los seres humanos tienen conocimiento del uso de la lengua, pero que sus procesos mentales estructuran la manera en que se utiliza. Por lo tanto, podemos suponer que cada persona tendrá diferencias idiosincráticas con respecto a los demás.

Con base en los postulados teóricos sobre las transiciones de palabras, se aborda de manera cuantitativa el análisis de secuencias mediante el uso de bigramas de las variables de estudio. Estos componentes representan combinaciones específicas de sílabas, con el objetivo de examinar las frecuencias relativas en un corpus de estudio y determinar hasta qué punto constituyen marcas distintivas.

3. Comparación de hablantes y atribución de autoría

El lenguaje es el resultado del contacto del individuo con el entorno y es adquirido según la interacción con los demás miembros de la comunidad lingüística a lo largo de su vida, estas particularidades individuales son las que constituyen el idiolecto de cada persona (Chomsky, 1957). Dichas particularidades pueden ser causadas ya sea por aspectos físicos, como la disposición y resonancia de los órganos fonoarticuladores, o por factores cognitivos, que incluyen el pensamiento, las creencias y experiencias del hablante. En el contexto forense y, específicamente, en el ámbito de la comparación de hablantes, se emplea como objeto de estudio las propiedades acústicas donde se busca los diferenciales fonéticos de cada individuo, así como el estudio de marcas distintivas en el lenguaje (Morrison et. el., 2019).

Con relación a los parámetros acústicos, los estudios buscan proporcionar un perfil mediante parámetros cuantificables, como la altura de los formantes o la inflexión tonal, logrando establecer marcas propias de cada sujeto de estudio. Ahora bien, en relación con la búsqueda de parámetros estilísticos, se evalúa las secuencias de bigramas de palabras considerándolas como elementos dependientes de las decisiones particulares en la selección y producción léxica según el uso lingüístico normativo con el ánimo de identificar patrones parciales de orden en las secuencias. Es entonces, cuando las emisiones recolectadas y su transición son utilizadas como elemento idiosincrático distintivo que refleja el uso del repertorio léxico acuñado por las experiencias y la relación social del hablante.

Debido a la estrecha relación entre el habla y la escritura, es posible explorar un enfoque en las frecuencias de las secuencias silábicas, de manera similar a cómo se abordan los estudios que analizan textos escritos con el objetivo de comparar elementos idiolectales en la atribución de autoría (Coutlhard, 2016). Este enfoque, se sustenta en la premisa que cada

usuario tiene su propia versión de lengua (Bloch, 1948) respaldado por la variación lingüística, la cual implica que un individuo puede ser identificado a partir de su voz o de sus escritos. Estas son manifestaciones factibles del código lingüístico que se transmiten a través del lenguaje y presenta el enfoque idiolectal aplicado en las transcripciones de los hablantes (Stefanova, 2009).

Ahora bien, los sistemas de comparación forense automático de hablantes emplean diversos modelos de análisis como son MFCC, i-vector, x-vector, entre otros (San Segundo et. al, 2019) para la evaluación de los parámetros acústicos. En el caso de la atribución de autoría, Stamatatos (2009) propone cinco características estilométricas para el análisis del texto: característica léxica (frecuencia de palabras funcionales, riqueza del vocabulario, etc.), característica de caracteres (frecuencia de letras y n-gramas de caracteres), característica sintáctica (frecuencia de las etiquetas, mediciones en la estructura de oraciones y frases, etc.), característica semántica (mediciones de sinónimos, de dependencias semánticas, etc.) y características específicas de la aplicación (tamaño de fuente, color de la fuente, uso de tecnolecto, etc.). Velásquez et al. (2019) han encontrado que las mediciones más efectivas para la atribución de autoría son las características léxicas.

Existen distintos análisis que se pueden ejecutar bajo la estilometría, entre las cuales se encuentra la atribución, verificación o perfilación de autoría, la estilocronometría y la estilometría adversarial (Neal et. al., 2017). El uso de la estilometría como herramienta de estudio ha sido ampliamente utilizada sobre textos escritos analizando aspectos como el uso de las puntuaciones, la longitud de las oraciones, el uso de las mayúsculas, entre otros. Sin embargo, el uso de estilometría en la comparación de hablantes utiliza otros elementos de estudio debido a que las transcripciones de audios carecen de elementos estilísticos

tradicionales por lo que es necesario enfocar la evaluación en otras características tales como las combinaciones de las sílabas y su posición en la cadena hablada, la riqueza del vocabulario, entre otros.

3.1. Establecimiento de la selección de sílabas como marcas distintivas

Este estudio busca establecer marcadas identificativas fundamentadas en la selección y frecuencia silábica de los hablantes, destacando su relevancia en la configuración estilística individual. A través del establecimiento de los valores idiosincráticos, se busca demostrar que el uso de las transiciones silábicas es un elemento propio de cada sujeto.

Como primera medida es requerido establecer si según las bases de los modelos biométricos, las sílabas son elementos distintivos. Según Maltoni (2003), la variable debe cumplir algunos requisitos que determinan la aptitud de una característica biométrica:

- Universalidad: cada persona tiene una característica biométrica
- Exclusividad: cada persona es única, en términos de las características, lo cual hace que sea diferente de los demás.
- Perdurabilidad: las características biométricas no cambian con el tiempo.
- Cuantificable: las características biométricas pueden ser medidas.
- Rendimiento: la identificación debe ser ágil y precisa.
- Aceptación: los métodos tecnológicos deben ser aceptados por la comunidad científica.
- Elusión: el sistema no puede ser engañado fácilmente.

Según el autor, es difícil que un sistema biométrico cumpla con todas las premisas propuestas debido a que cada sistema tiene sus fortalezas y debilidades dependiendo de la naturaleza y los requisitos de aplicación de las características biométricas.

En el caso de las sílabas cumplen con las características de universalidad al ser, morfológicamente, la unidad de organización secuencial de los sonidos de habla y la menor división de la cadena hablada para la conformación del lenguaje. La exclusividad es la característica que se evalúa en el artículo bajo la premisa de que la selección y su organización en la emisión fraseológica es dependiente del pensamiento individual. Ahora bien, en términos de perdurabilidad el lexicón se va adaptando según el entorno y la influencia de factores externos, lo cual representa una limitante a nivel temporal del estudio.

Las sílabas como unidades pueden ser cuantificables y, por tanto, es posible tener una caracterización numérica que permite observar la organización y selección silábica para poder ser estudiada de manera ágil y con el menor costo computacional posible. Para ello, se establecen parámetros que funcionen de manera adecuada que solamente sean modificados según la extensión de los conjuntos de datos.

Debido a que el factor evaluado corresponde a una característica cognitiva que mide la eficiencia de cómo se procesan las sílabas en el cerebro, base para la identificación biométrica basada en el lenguaje (Bhattacharyya et. al., 2009), el sistema no puede ser engañado fácilmente al ser procesos inconscientes.

4. Trabajos relacionados

El análisis de comparación de hablante en ambiente forense está orientado casi en su totalidad a la evaluación de los parámetros acústicos y fonéticos. Sin embargo, el estudio desarrollado por Sztahó et al. (2022) utiliza el modelado vectorial de párrafos sobre las transcripciones de muestras de habla espontánea para luego, por medio de regresión logística de la distancia coseno, predecir la probabilidad de que sea o no el mismo autor. El conjunto de datos utilizado es ForVoice120+ (Beke, 2021), el cual registra muestras de habla espontánea

de 120 hablantes con tres diferentes estilos de habla: diálogo libre, entrevista guiada y monólogos. Para el entrenamiento del algoritmo utiliza dos bases de datos adicionales: BEA (Gósy 2012) y HuComTech (Szekrényes, 2014) para un total de 182 hablantes que son utilizados para el entrenamiento y 20 para la prueba.

El estudio desarrollado por Sztahó tiene similitudes con el método propuesto dado que su motivación es el uso de Doc2VeC, algoritmo propio de clasificación de autores y perfilación de autor en textos escritos (Le et al. 2014), obteniendo vectores de tamaño 20, 100 y 200 para sus experimentos para obtener una representación vectorial en la evaluación de evidencias por medio del índice de verosimilitud.

La evaluación del rendimiento del modelo propuesto es obtenida con el Equal Error Rate (EER). Los resultados variaron entre 0.35 y 0.11. Aunque son resultados relativamente modestos, demuestran que el texto transcrito puede contribuir para la verificación de hablantes. Sin embargo, un error del 35% indica que puede haber muchos casos de falsa aceptación o falso rechazo lo cual perjudica la fiabilidad de la comparación.

El estudio tiene algunas debilidades debido a que no expone la forma en que se realizó la segmentación de los párrafos de entrenamiento y de evaluación. Además, no es posible determinar si las decisiones tienen un mismo modelo semántico que puede limitar y sobreentrenar el modelo. Así, es posible que los resultados obtenidos dependieran de la temática que tratan los individuos en las muestras, los cuales proveen una identificación débil en comparación con identificaciones más robustas.

En 2023, Sztahó realizó un nuevo acercamiento al uso del corpus de estudio, pero esta vez no utilizando las transcripciones sino los audios. En este caso, utilizó deep learning sobre vectores embebidos que representaban a los hablantes utilizados para el entrenamiento y la

prueba del modelo. El estudio demostró una mejora significativa en la métrica EER ya que los resultados ahora estaban en el rango entre 5% y 11% usando muestras de 2 segundos para la prueba del modelo.

Ahora bien, Abed (2023) hizo pruebas de identificación utilizando redes neuronales que combinan elementos convolucionales (CNN) con una variante de las redes recurrentes (RNN) conocida como *Time Delay Neural Network.* Este sistema de ECAPA-TDNN utiliza conexiones con retrasos temporales fijos para capturar patrones en los datos secuenciales y permiten a la red procesar datos en diferentes puntos temporales y aprender características en los datos de entrada incorporando la información contextual en el modelo. Esto significa que la red no solo tiene en cuenta la entrada actual, sino también información contextual de la secuencia completa de datos. Los resultados obtenidos con esta arquitectura son cercanos al 2% en húngaro e inglés.

5. Método

Con la concepción de que las personas pueden ser identificadas por medio de alguna característica fisiológica según parámetros base de comparación (Bhattacharyya, 2009). En este estudio se usa la secuencia y frecuencia silábica, para generar vectores multidimensionales para pares de muestras de *clase 0* (muestras extraídas de un mismo hablante) y *clase 1* (pares de muestras de comparación que contienen información de un mismo hablante y de un hablante diferente). Posteriormente, estas son comparadas en volumen con Aprendizaje Maquinal (*Machine Learning)* para obtener un modelo que pronostique la similitud y clase entre nuevos pares de muestras “no vistos” en el proceso del modelo. Finalmente, estas predicciones son evaluadas para obtener el índice de error de la predicción del algoritmo.

A continuación, presentamos una visión general del método utilizado, la cual se detalla en las subsecciones posteriores. Para la construcción y evaluación del modelo de verificación se utilizan muestras recolectadas bajo los criterios de Morrison et al. (2012) para la simulación de ambientes forenses. En este caso, uno de los *dataset* utilizados está en húngaro y consta de 120 hablantes que participan en dos sesiones separadas por dos semanas. Cada sesión contiene muestras de habla espontánea con tres estilos de habla distintos: diálogo libre (aprox. 10 minutos), diálogo orientado (aprox. 8 minutos) y monólogos (aprox. 3 minutos). La transcripción y corrección de los corpus fue realizada manualmente. El segundo *dataset* utilizado está en inglés y consta de 3,899 grabaciones que tienen, en total, una duración de 310 horas. La construcción de la base de datos se hizo con 555 hablantes. Debido a la gran diferencia de hablantes entre los dos conjuntos, se seleccionaron únicamente las muestras de los hablantes masculinos que participaron en al menos dos sesiones, dando como resultado 167 sujetos de estudio.

Al igual que la base de datos en húngaro, la separación entre sesiones fue de 2 semanas. Sin embargo, para este caso las tareas de cada sesión consistían en una conversación telefónica casual, intercambio de información para completar una tarea por el teléfono y una entrevista policial. Más información sobre los conjuntos de datos utilizados se presenta en la subsección 5.1.

A los dos conjuntos de datos se les realiza una segmentación en sílabas para su análisis. Debido a que los conjuntos relativamente pequeños en su extensión, usamos el método propuesto por Sztahó y Fejes (2023), quienes traslapan cada muestra según un desfase (T) determinado por la duración del corpus. Con las muestras obtenidas se construyen los vectores de comparación según las frecuencias relativas en un corpus de referencia los cuales

se agrupan en *clase 0* o *clase 1*. Los vectores matriciales se dividen en dos grupos: entrenamiento y prueba (extraído de la *clase 0*), los cuales son usados por el modelo computacional en la determinación de las diferencias y similitudes en el uso de las frecuencias de las transiciones de sílabas. Finalmente, se evalúa el rendimiento con EER.

Una de las diferencias entre modelo propuesto y los modelos tradicionales basados en texto es, que estos últimos analizan características estilométricas léxicas para la atribución de autoría como el análisis de palabras funcionales (García et al., 2006) o el uso de marcas sintácticas (Martinic, 2017) entre los mismos textos de comparación. En su lugar, se analizan las transiciones de sílabas indexadas por frecuencias a partir de un corpus de referencia. Para ambas lenguas se utilizó una base de datos que contenía ya la información extraída de la frecuencias de palabras en Twitter, periódicos y blogs llamada Wordlex[2], infiriendo las frecuencias de sílabas a partir de frecuencias de palabras para la representación numérica de las sílabas (Gimenes & New, 2015).

Para el modelo, establecemos como "muestra dubitada" m_d las locuciones registradas en la sesión 1 y para la "muestra indubitada" m_i, la sesión 2. Buscando la verificación de un hablante, extrae aleatoriamente m_d del conjunto de datos y se procede a asociar las otras muestras generadas por ese mismo hablante objetivo (y otros de contraste) en las sesiones 1 y 2 (*clase 0*) y, dependiendo de las combinaciones posibles, se procede a obtener un número igual de muestras escogidas al azar con los demás hablantes de contraste (*clase 1*). Luego, estos conjuntos de pares de muestras clasificados son utilizados para construir un clasificador por medio de *machine learning* . Luego, este clasificador es probado con la m_d pareada con otras muestras del hablante objetivo (clase 0) y con muestras de hablantes de contraste (clase

[2] http://www.lexique.org/?page_id=250

1). El rendimiento del clasificador en esta prueba dará indicios de la efectividad de la frecuencia y secuencia silábica en la tarea, y que tan distintivo es esta representación para la verificación de los hablantes.

5.1. Datos

Para simular el escenario forense se utilizaron dos conjuntos de datos que contienen grabaciones en ambientes controlados siguiendo el protocolo descrito por Morrison et al. (2012). Los conjuntos se denominan AusEng500 (Morrison et al, 2021) en inglés y ForVoice120+ para el húngaro.

El corpus AusEng500 tiene más de 500 hablantes que van desde los 18 hasta los 70 años. Cuenta con 231 grabaciones a hombres de las cuales 62 solamente estuvieron presentes en una sesión, 43 en dos sesiones, 107 en tres sesiones y 19 en más de tres sesiones. Las tareas para las sesiones fueron: conversación casual telefónica, intercambio de información a través del teléfono y una entrevista con un policía. Se seleccionaron las muestras de sujetos que estuvieran presentes en al menos dos sesiones[3]. Las grabaciones de audio fueron transcritas con la herramienta *Whisper* (Radford et al., 2023), la cual está disponible para uso gratuito y es uno de los mejores sistemas de reconocimiento automático del habla (*ASR*, por su sigla en inglés) debido, principalmente, a su entrenamiento masivo con más de 680.000 horas de audio etiquetado (Gong et al., 2023). En total, resultaron 167 locutores con un total de 908.222 palabras. Ahora bien, para la comparación se necesita que sean 2 muestras a comparar separadas en el tiempo, motivo por el cual, se utilizó la sesión 1 contra las demás sesiones juntas.

[3] Fueron descartados dos sujetos debido a que la transcripción realizada tuvo problemas con la silabación.

El corpus ForVoice120+ contiene las muestras grabadas de 59 hombres y 61 mujeres con edades entre 18 y 50 años en dos sesiones de grabación. Ambas sesiones de grabación se dividen en tres tareas. La primera tarea corresponde a diálogo libre en donde la persona realiza una comunicación con alguien más y se discuten temas generales, se simula, por tanto, una interceptación telefónica. La segunda tarea consiste en un intercambio de información, la persona en este caso simula estar planificando ejecutar una actividad delictiva y la tercera tarea consiste en un monólogo donde cuenta aspectos generales de su vida. En total, se utilizaron 320.307 palabras. El corpus fue transcrito y corregido de manera manual por quienes realizaron las grabaciones de las sesiones.

5.2 Extracción de unidades silábicas

Una vez obtenidos los corpus de estudio se procedió a realizar la división silábica, proceso que se realizó con la librería pyhyphen[4], la cual es un algoritmo que está basado en reglas lingüísticas de las lenguas que lo componen, las cuales incluyen el inglés y el húngaro entre otras 45 más. Una de las motivaciones para el uso de sílabas en lugar de las palabras completas es que se logra aumentar el tamaño de las muestras para realizar las comparaciones. Por ejemplo, en la palabra *computer*, se podría representar en tres unidades *com-pu-ter*, en lugar de una sola. Así, se logra aliviar el problema léxico debido a que las muestras que llegan a los laboratorios periciales pueden tener una limitada duración, sin embargo, depende de la riqueza lingüística para el estudio. Para el corpus AusEng500 se obtuvo 1,056.478 sílabas con un promedio de 6.326 sílabas por hablante y 1,16 sílabas por palabra y para el corpus

[4] Para la silabación de los datos tanto en inglés como en húngaro se utilizó el paquete PyHyphen disponible en https://pypi.org/project/PyHyphen/

ForVoice120+ se obtuvo 574,173 sílabas con un promedio de 4,784 sílabas por locutor y 1.79 sílabas por palabra.

5.3. Preprocesamiento de los datos

Una vez recolectadas las muestras, y obtenida la división silábica de las transcripciones de los audios, se obtiene representación vectorial basada en las transiciones de sílabas y frecuencias de aparición en el *WordLex*. Luego, esas representaciones vectoriales son comparadas en pares donde cada vector corresponde ya sea a una muestra dubitada o a una muestra indubitada. Cada uno de estos pares puede corresponder a la *clase 0*[5] o a la *clase 1*. Se busca obtener un número considerable y balanceado de pares de ambas clases para el entrenamiento y la prueba del modelo. Estas listas de pares, luego, son utilizadas en combinación con la regresión logística para su clasificación y análisis. Así las cosas, se definen los siguientes parámetros que son utilizados para el procesamiento:

m_d	Tamaño de las muestras dubitadas extraídas de la sesión 1.
m_i	Tamaño de las muestras indubitadas extraídas de las demás sesiones.
T	Desfase de las muestras (ver Sztahó, 2022).
G	Grilla, número de segmentos en los que se divide el espectro de frecuencias de sílabas.
θ	Umbral mínimo de apariciones de una transición de sílabas para ser considerada.

[5] Hipótesis nula de la Fiscalía que establece que la muestra dubitada fue producida por el hablante objetivo o indiciado.

5.3.1 Indexamiento por frecuencia.

La representación vectorial para el lenguaje consiste en convertir secuencia de símbolos (i.e. palabras o sílabas) de longitud variable en una lista de números configurados en un vector de tamaño fijo. La representación más usada es donde cada dimensión del vector indexa una palabra del vocabulario y su contenido es el conteo de apariciones de cada palabra en la secuencia (Salton et al., 1975). Sin embargo, en este caso el largo del vector es del orden del tamaño del vocabulario generando una representación dispersa, donde la mayoría de las entradas del vector son ceros. El opuesto es una representación densa donde el vector sea "corto" (del orden de centenas) y con pocas entradas nulas. Aunque las representaciones densas son útiles para construir sistemas de aprendizaje automático, usualmente son de baja interpretabilidad en comparación a las dispersas (Faruqui et al, 2015). En este trabajo, usaremos una representación vectorial indexada por frecuencias inspirada en las Firmas Espectrales de Corpus (*CSS*) de Jimenez et al. (2020), las cuales son densas pero interpretables. Así, el rango de frecuencias de las sílabas en el corpus de referencia Wordlex, es primero transformado logarítmicamente y luego dividido en G partes iguales, donde el primer segmento contiene las frecuencias más bajas y el último las más altas. En la práctica, las muestras de habla en audio transcritas y segmentadas en sílabas, son transformadas de una secuencia de sílabas a una secuencia de índices de las G partes del espectro de frecuencias segmentado. Esto reduce el número de símbolos posibles en la secuencia desde el tamaño del vocabulario de sílabas, que es del orden de miles, a solo G símbolos del orden de decenas. Dado que el espectro de frecuencias en Wordlex es amplio, su división en solo G pares es una división "gruesa". Por tal motivo denominamos el parámetro G como "grilla".

El corpus de contraste utilizado para obtener las frecuencias de las sílabas fue creado por Gimenes y New (2016) quienes recopilaron las tablas de frecuencias de palabras para 64 lenguas desde páginas web como blogs, Twitter y periódicos. Estas tablas son utilizadas para establecer la frecuencia de cada una de las palabras de las sesiones y, de esta manera, inferir las frecuencias de las sílabas.

La transformación logarítmica está motivada tomando como base la Ley de Zipf, haciendo que cada una de las G entradas en la grilla correspondan a un número aproximadamente uniforme de sílabas.

5.3.2. Segmentación en muestras.

Los tamaños de muestras m_d y m_i hacen referencia al número de sílabas extraídas de las transcripciones y su tamaño está limitado por el número de sílabas disponibles en las muestras de audio de las sesiones. Ahora bien, el procedimiento requiere que exista una cantidad adecuada de muestras de secuencias de sílabas para cada hablante, las cuales se extraen teniendo en cuenta el traslapo T y, de esta manera, obtener una mayor cantidad de muestras para el entrenamiento y prueba del clasificador. Este método fue utilizado por Sztahó y Fejes (2022) para aumentar el número de muestras disponibles por hablante a partir del audio de las sesiones.

Una vez indexadas todas las sílabas en los datos en la grilla de tamaño G, se extrae y remueve de manera aleatoria una "muestra dubitada" de la sesión 1 para el hablante objetivo. Luego, para cada hablante en el conjunto de datos (incluido el hablante objetivo) se realiza la división de la sesión 1 en muestras de largo m_d y de la sesión 2 en muestras de tamaño m_i, ambas con traslapo T. Con el fin de que el clasificador "aprenda" un modelo ajustado al hablante objetivo, el traslapo de este es reducido a $T/10$ aumentando así su número de

muestras. El agrupamiento de las sílabas en las muestras dubitadas (sesión 1) y las muestras indubitadas (sesiones 2 y 3). Ahora bien, el traslapo es dependiente del tamaño del corpus motivo por el cual AusEng500 tiene un valor T de 140 mientras que el de ForVoice120+ es de 100, haciendo así que la cantidad de muestras en ambos conjuntos de datos sea comparable.

El resultado de este proceso es una muestra dubitada para un hablante objetivo, y para cada hablante de contraste (incluyendo al hablante objetivo) un conjunto de muestras de tamaño m_d proveniente de la sesión 1, y otro conjunto de muestras de tamaño $m_i$de la sesión 2.

5.3.3. Representación vectorial por transiciones de frecuencias.

Para codificar en la representación vectorial de las muestras una noción de orden de las secuencias utilizamos las transiciones o bigramas. De esta forma, la representación vectorial tiene un tamaño de $G \times G$ donde las entradas en los vectores indexan transiciones de un segmento de frecuencias de sílabas a otro y los valores son los conteos del número de apariciones de esa transición en cada muestra. Estos conteos se transformaron logarítmicamente siguiendo esta práctica usual en el campo del Procesamiento del Lenguaje Natural (Lan et al., 2005). Así, la presentación captura un modelo de orden parcial Markoviano de primer orden, o sea solo transiciones.

Sin embargo, existe una gran cantidad de transiciones de segmentos de frecuencias silábicas que son poco frecuentes. Entonces, se hace necesario establecer un filtro a la cantidad de apariciones de transiciones entre rangos de frecuencias θ, que para el caso de ForVoice120+ se determinó en al menos $\theta = 10$ apariciones que sean relevantes y en AusEng500, debido a que es aproximadamente tres veces más grande la cantidad de sílabas, se estableció en $\theta = 30$. De esta forma se logra reducir el número de dimensiones del vector.

Por ejemplo, para AusEng500 usando $G = 30$, se redujo la dimensionalidad del vector de 900 transiciones posibles a 346, logrando así una representación más densa sin afectar la interpretabilidad del indexamiento.

Finalmente, esta representación modela las posibles diferentes zonas del silabario mental de un hablante simplificado en los G rangos de frecuencias, y las transiciones modelan las posibles interconexiones entre las zonas del silabario mental. El resultado final de este proceso es que cada una de las muestras obtenidas en la subsección anterior se transforman en vectores de alrededor de 400 dimensiones. La dimensionalidad exacta depende principalmente de G^2 y del número de transiciones de rangos de frecuencias que superen el umbral θ.

La Figura 1 ilustra el proceso de preprocesamiento de los datos descrito en las subsecciones 5.3.1, 5.3.2 y 5.3.3.

5.3.4 Asociación de las muestras representadas como vectores.

Una vez decantada la información y obtenidas las representaciones vectoriales de las muestras producidas por cada uno de los hablantes, se realizan las asociaciones de los vectores para construir un modelo de clasificación que permite verificar si una muestra dubitada fue producida o no por un hablante objetivo, la finalidad consiste en construir a partir de los vectores dos conjuntos de datos, uno para entrenamiento del modelo de clasificación y otro para prueba.

Para iniciar, contamos con una muestra dubitada de un hablante objetivo, la cual fue retirada de los datos y para cada hablante se dispone de un conjunto de vectores proveniente de las muestras de tamaño m_d sílabas de la sesión 1 (conjunto C_d), y otro conjunto de vectores proveniente de las muestras de tamaño m_i sílabas de la sesión 2 (conjunto C_i). El proceso de construcción del conjunto de entrenamiento es el siguiente. Primero, para cada hablante

asociamos en pares ordenados todas las muestras de C_d y C_i obteniendo así una cantidad de $|C_d \times C_i|$ pares de vectores los cuales, tomamos para cada hablante los vectores de su C_d y los asociamos con muestras tomadas aleatoriamente de hablantes diferentes de sus C_i cuidando que para cada hablante se obtengan el mismo número de pares que se obtuvieron en la clase 0. De esta manera se obtiene un conjunto de datos de pares de vectores balanceado entre sus clases y de tamaño máximo. Ahora, para que este conjunto de entrenamiento sea usable por cualquier algoritmo de clasificación, es necesario convertir cada par de vectores en un solo vector. Para esto usamos el método tradicional de usar el producto Hamdard (Islam et al, 2023), el cual consiste en multiplicar en parejas las entradas correspondientes de los dos vectores. Todos los vectores de entrenamiento se organizan en una matriz la cual denominamos $\mathbf{X}_{train}$ y sus clases correspondientes en un vector columna denominado $\mathbf{Y}_{train}$. Es importante tener en cuenta dos factores importantes en este conjunto de entrenamiento: primero, el hablante objetivo está "sobre-muestreado" por el uso de un traslapo diferencial de $T/10$ haciendo que aproximadamente el 20% de los ejemplos de entrenamiento involucra una muestra del hablante objetivo; y segundo, dado que la primera acción es remover la muestra dubitada de los datos, en la matriz $\mathbf{X}_{train}$ no existe información alguna de esta muestra. Estos factores hacen que el modelo resultante modele principalmente al hablante objetivo pero que no se use información alguna de la muestra dubitada.

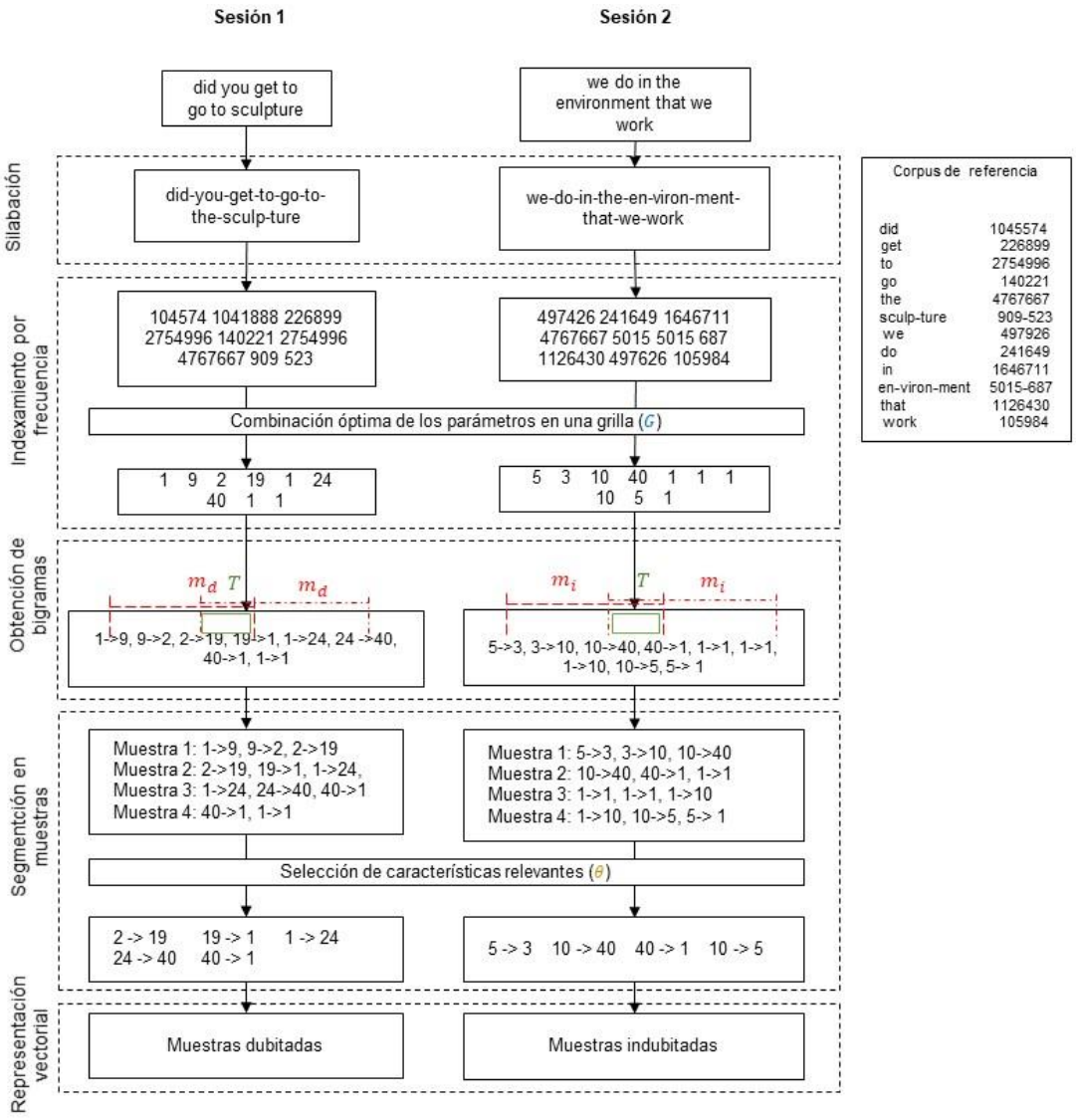


Ilustración 1. Diagrama de flujo del indexamiento por sílabas

Para construir el conjunto de prueba, primero se asocia el vector de la muestra dubitada con todos los vectores C_i del hablante objetivo, y se etiquetan con la clase 0 (mismo hablante objetivo). Para obtener los pares de vectores de la clase1 (hablante objetivo vs. otros hablantes) se hacen pares con el vector de la muestra dubitada y vectores seleccionados aleatoriamente de los conjuntos C_i correspondientes a otros hablantes diferentes al hablante objetivo. Al igual que para el conjunto de entrenamiento los pares de vectores se fusionan con usando el producto Hamdard obteniendo la matriz $\mathbf{X}_{test}$ y el vector columna $\mathbf{Y}_{test}$ .Es

importante anotar que $\mathbf{X}_{train}$ contiene considerablemente muchos más ejemplos (cientos de miles) que los que hay en $\mathbf{X}_{test}$ (centenas). Esto se debe a que en la construcción del modelo se utiliza toda la información posible disponible del hablante objetivo y de los hablantes de contraste para lograr el objetivo de verificar solo una muestra dubitada, lo cual es concordante con el escenario forense real. El modelo a construir usando $\mathbf{X}_{train}$ y $\mathbf{Y}_{train}$ será evaluado luego con $\mathbf{X}_{test}$ y sus predicciones deberían coincidir con $\mathbf{Y}_{test}$. La Figura 2 muestra un esquema de la configuración de los datos para entrenamiento y prueba o validación.

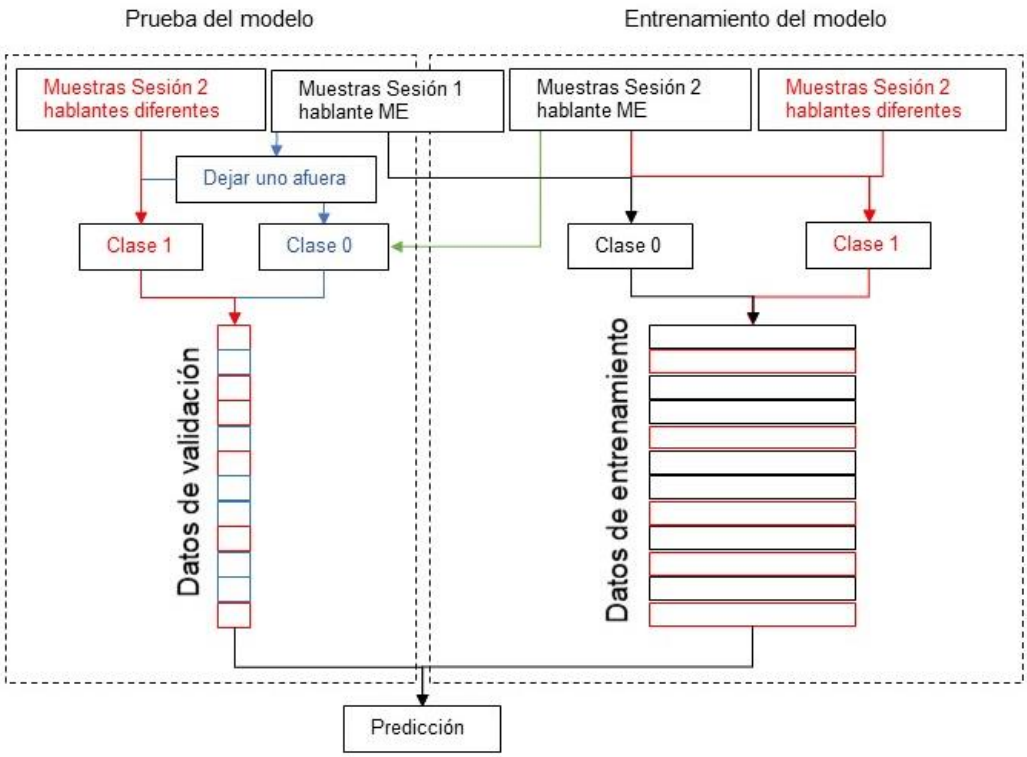


Ilustración 2. División de los datos en muestra dubitada, entrenamiento y prueba.

5.4. Sistema de verificación de hablantes basado en regresión logística

La Regresión Logística es un análisis estadístico para clasificación en dos clases que predice el resultado de una variable dependiente (i.e. clase 0 o 1) en función de vectores que contienen las variables independientes (i.e. los conteos de rangos de frecuencias silábicas

(LaValley, 2008). Similar a la Regresión Lineal, la Regresión Logística provee un conjunto de coeficientes de "importancia" asociados a cada una de las variables independientes representando en qué medida la variable dependiente contribuye a la clase 0 (coeficientes negativos) o a la clase 1 (positivos). Esta cualidad de interpretabilidad es ideal para preservar la interpretabilidad de representación vectorial basada en transiciones de rangos de frecuencias silábicas. Los coeficientes $\beta_1, \beta_2, \ldots, \beta_n$, uno por cada entrada en los vectores de entrenamiento, indican la magnitud de la contribución de cada característica indexada en las entradas de los vectores a la clasificación binaria. Las predicciones para cada vector es la probabilidad de que las dos muestras de habla que produjeron el vector pertenezcan a dos hablantes diferentes. De manera complementaria, si esta probabilidad es cercana a cero significa que las dos muestras que produjeron el vector pertenecen al mismo hablante. Esta probabilidad para un vector $\boldsymbol{x}$ en la matriz de entrenamiento $\mathbf{X}_{train}$ se calcula así:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Los coeficientes β se obtienen minimizando la medida de entropía cruzada[6] entre las predicciones y los valores reales de la variable dependiente en $\mathbf{Y}_{train}$. Luego de "entrenado" el modelo se obtienen predicciones para los vectores y estas se comparan contra los valores reales de la variable dependiente. En principio, si el modelo logra "aprender" del conjunto de datos de entrenamiento, las predicciones serán cercanas a 0 para los vectores de la clase 0 y a 1 para los de la clase 1.

[6] Para esto utilizamos la implementación para Python disponible en https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

5.5. Obtención de firmas de transiciones espectrales silábicas

Dada la interpretabilidad de las entradas de los vectores (i.e. transiciones de rangos de frecuencias silábicas) y la interpretabilidad de sus coeficientes β asociados (i.e. contribución positiva o negativa de la transición a la clasificación), proponemos una visualización del modelo con el fin de evaluar la posibilidad de que en esta se distinga una "firma" o "huella" espectral visible que identifiquen y distingan a los hablantes.

Primero, el conjunto de datos obtenido $\mathbf{X}_{train}$ se somete a un proceso de filtrado con el propósito de asegurar que todas las instancias provengan de un par de muestras, donde al menos una de ellas sea del hablante objetivo. Este enfoque garantiza que la firma espectral represente al hablante objetivo en cuestión, al tiempo que aún incorpore información de otros hablantes de referencia.

En los conjuntos de datos utilizados para la verificación, el hablante objetivo está sobrerrepresentado, constituyendo aproximadamente el 20% de las instancias del conjunto. El restante 80% corresponde a pares de muestras con otros hablantes de contraste. Para mejorar la claridad de la firma espectral, nos enfocamos en el 20% que involucra exclusivamente al hablante objetivo. Los valores para la representación espectral se derivan de los coeficientes de la regresión logística, los cuales señalan las transiciones más importantes entre los grupos de frecuencias de las sílabas para distinguir entre muestras del hablante objetivo y las de contraste. Consideramos que estas importancias conforman la firma espectral característica del hablante objetivo, la cual debería ser consistente para un hablante objetivo cuando se obtiene de diversas muestras indubitadas del mismo hablante, y distinguirse de las firmas espectrales de otros hablantes.

La visualización de los datos depende de la cantidad de datos utilizados: con una gran cantidad de datos es más difícil observar diferencias en las firmas espectrales. Por tanto, configuramos la visualización de los datos a los 50 coeficientes β más relevantes (i.e. de mayor magnitud absoluta) para que los píxeles de la gráfica sean interpretables y comparables a simple vista. En la Figura 3 se muestra un esquema de la visualización propuesta. La idea es codificar en una matriz bidimensional las transiciones entre los rangos de frecuencias mostrando en la horizontal la primera sílaba de la transición y en la vertical la segunda. Tanto el eje horizontal como el vertical representa el espectro de frecuencias silábicas transformado logarítmicamente y segmentado en un número de G partes. En cada uno de las entradas o "pixeles" de la matriz se codifica el valor β obtenido del modelo de Regresión Logística codificando en color rojo los coeficientes positivos y en azul los negativos. La intensidad del color codifica la magnitud absoluta del coeficiente.

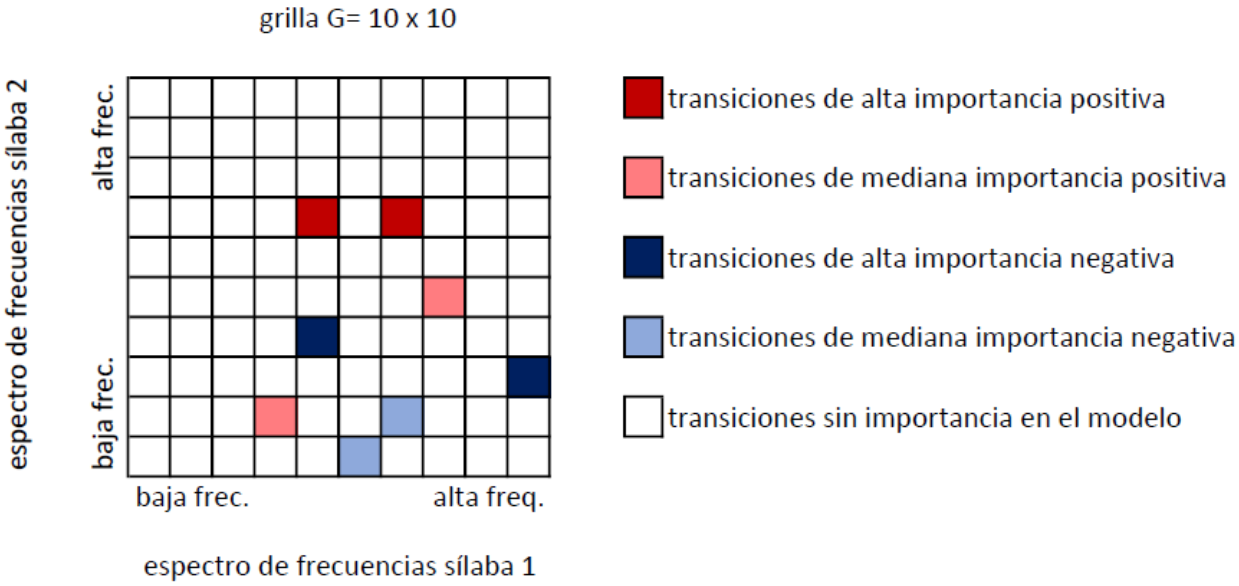


Ilustración 3. Modelo visual de una firma de transiciones espectrales silábicas.

Se espera que esta visualización refleje las posibles interconexiones más relevantes de las diferentes zonas del silabario mental almacenado cognitivamente por frecuencia para un

hablante. Si estas interconexiones relevantes son discriminantes entre los hablantes, es posible que surjan patrones claramente observables que diferencien un hablante de otro.

5.6. Arreglo experimental

5.6.1. Medida de rendimiento.

Con los conjuntos obtenidos se realiza la evaluación de la efectividad de las predicciones calculadas por el modelo con regresión lineal. Para la medición de la confiabilidad del método se utilizó el EER (*Equal Error Rate*), la cual es la medida más utilizada en la ciencia forense y en la identificación biométrica (Rose et al., 2009). Para utilizar esta medida se toman todas las predicciones del modelo de clasificación, las cuales son números reales entre 0 y 1, donde el 0 representa certeza total de que las dos muestras provienen de un mismo hablante y 1 lo contrario. Estas predicciones se ordenan numéricamente y se identifica el valor del umbral donde la tasa de falsos positivos es igual a la rata de falsos negativos, que es el valor de EER.

Ahora bien, los sistemas de verificación utilizan, además del EER, el Cllr (*Cost of log-likelihood ratio*) como medida de rendimiento. Esta es una función que mide el balance de los puntajes de LR (likelihood ratio) de las muestras comparadas que tienen el mismo origen y las de origen diferente. Para el caso que se presenta, se decide únicamente utilizar el EER debido a que la correlación de Pearson entre ambos resultados es de 0.987 según los resultados de la Figura 8 en el estudio de Sztahó & Fejes (2023). Por esta razón y para ahorrar espacio en la

sección de resultados adoptamos como medida de rendimiento solamente a EER y la consideramos en la práctica equivalente a Cllr[7].

5.6.2. Validación dejando "una muestra fuera".

El modelo presentado permite la verificación de si una muestra dubitada simulada corresponde o no a un hablante objetivo. Esto se logra extrayendo y removiendo esta muestra dubitada simulada de las muestras disponibles del hablante objetivo. Una vez removida la muestra, el método propuesto utiliza toda la información del resto del conjunto de datos para construir un modelo predictivo de clasificación para el hablante objetivo. Este modelo se evalúa pareando la muestra dubitada con otras muestras del hablante objetivo (clase 0) y con otras muestras de otros hablantes (clase 1). Así, usando EER, cuantificamos el rendimiento en la verificación de la muestra dubitada. En los sistemas de identificación, un valor EER cercano a cero significa que el sistema tiene una alta certeza en la tarea de identificación o verificación.

Esto nos permite obtener un valor de EER para una muestra dubitada de un hablante objetivo. Sin embargo, dado que la muestra dubitada fué escogida al azar, otra muestra dubitada del mismo hablante obtendrá un valor de EER diferente. Para tener en cuenta esta posible variación, repetimos el proceso de seleccionar aleatoriamente la muestra dubitada 10 veces y reportamos para cada hablante el promedio y desviación estándar de los 10 valores de EER obtenidos por hablante. Igualmente, para reportar un valor único de EER para cada uno de los dos conjuntos de datos usados (AusEng500 y ForVoice120+) se promedian los promedios obtenidos por cada uno de los hablantes.

[7] Cuaderno de Colab con la comparación de entre EER y Cllr, y la obtención del modelo lineal para estimar Cllr a partir de EER: https://colab.research.google.com/drive/19VJOsudi3z7Wfr5j3jPoPSVe6NcuWmvD?usp=sharing

Ahora bien, en el caso del proyecto realizado por Sztahó, se utilizaron 40 hablantes del ForVoice120+ para el entrenamiento y los 80 restantes para pruebas. Debido a que son pocos los hablantes del entrenamiento utilizaron, además, dos bases de datos que no tienen propósito forense obteniendo un total de 632 hablantes para el entrenamiento. En el caso del conjunto AusEng500 en inglés, fueron utilizados 395 hablantes para entrenamiento, 80 para calibración del sistema y 80 para evaluación. En comparación, el modelo de evaluación de "una muestra fuera" se asemeja al escenario forense donde el objetivo es verificar la autoría de una muestra particular, en lugar de construir un sistema para identificar 40 u 80 hablantes simultáneamente.

Además de los corpus mencionados, los autores mencionados utilizan el esquema de sistema de identificación como en VoxCeleb, donde se busca identificar un conjunto de hablantes con un sistema entrenado con sus muestras. Para ello se dividen los hablantes entre las tareas de entrenamiento y de prueba para luego utilizar validación cruzada. Aunque este tipo de procedimiento no coincide con el ambiente forense, se utiliza para verificar la eficiencia del método propuesto. En el escenario forense se tiene una muestra dubitada recolectada ya sea por medio de interceptación de comunicaciones o por grabación desde un dispositivo encubierto donde participa una persona de quien se presume que está cometiendo un acto punible y se requiere cotejar esa muestra recolectada, generalmente con condiciones adversas, con muestras tomadas en el laboratorio con condiciones ideales.

5.7. Exploración del espacio de parámetros

Para obtener los resultados del sistema de verificación de hablantes, conlleva un alto costo computacional y de tiempo de procesamiento debido a la gran cantidad de datos que debe comparar. Por lo tanto, hacer una exploración factorial del espacio de las posibles

combinaciones de los valores de los parámetros no es, económicamente, factible. Así las cosas, escogimos un conjunto de valores plausibles para los parámetros y a partir de este hacemos variaciones de cada parámetro a la vez para reportar las variaciones en el rendimiento del sistema. Estos valores por defecto son los siguientes:

Tabla 1

Valores de los parámetros definidos por defecto.

Parámetro	**Inglés**	**Húngaro**
Tamaño muestra dubitada (m_d)	300	300
Tamaño muestra indubitada (m_i)	700	700
Traslapo de las muestras (T)	140	100
Grilla de división de transiciones (G)	40	40
Umbral mínimo de frecuencias (θ)	30	10

El tamaño de la muestra dubitada fue escogida teniendo como base las muestras de palabras elegidas por Sztahó (2022) quien utilizó 200 palabras por muestra con un desfase de 100 palabras en cada muestra. Sin embargo, como el sistema trabaja con sílabas se aumenta a 300 el tamaño de la muestra dubitada pero el parámetro del desfase se deja sin modificar. Así mismo, como la muestra indubitada generalmente es recolectada con condiciones ideales en un laboratorio y depende es del experto que realiza la toma de muestra de habla, el resultado da una más grande y, por tanto, el tamaño de extracción de sílabas es mayor. Es por ello que se selecciona el tamaño de 700 por defecto para las muestras indubitadas.

Ahora bien, el tamaño del corpus en inglés es aproximadamente tres veces el tamaño del húngaro, motivo por el cual es posible aumentar el tamaño del desfase de las sílabas y de esa manera disminuir la cantidad posible de muestras y no saturar el algoritmo con

demasiadas muestras, lo cual sobrepasa el límite de 12GB de las computadoras donde ejecutamos los experimentos. Seguidamente, se ubica un tamaño de grilla de agrupamiento que permite reducir la cantidad de datos para procesar, este valor es definido después probar los distintos valores con el objetivo de que las “firmas espectrales” fuesen interpretables debido a que a una mayor resolución es más difícil de apreciar los patrones visualmente. De la misma manera, para la categorización del umbral mínimo de frecuencias se configura un mayor tamaño para inglés comparado con húngaro ya que la frecuencia es directamente proporcional al tamaño del corpus. Es por ello que se ubica un umbral de 10 para el húngaro y de 30 para el inglés, correspondiente a aproximadamente el 1% de las transiciones en el corpus que es un filtro razonable y una práctica usual en el procesamiento de lenguaje natural donde se eliminan las apariciones de muy baja frecuencia.

6. Resultados

6.1. Análisis de la representación

Para el estudio estadístico de las transiciones de sílabas se utiliza el *análisis de componentes principales* (PCA por sus siglas en inglés). Este permite reducir la dimensión de las características y facilita la visualización de las transiciones de sílabas resaltando la relevancia de aquellos datos que no son correlacionados. A su vez, permite visualizar un conjunto de datos de alta dimensionalidad en 2D. Esta proyección sintetiza la información codificada en dos dimensiones y precisa de manera general la configuración y la agrupación de los datos para cada una de las clases. De acuerdo con el arreglo experimental, para cada muestra dubitada que será validada, se construye un conjunto de datos balanceado donde cada instancia es un par de muestras que corresponden a la *clase 0* o a la *clase 1*. Estas clases son utilizadas para las etapas de entrenamiento del algoritmo y su prueba. Al momento de evaluar

el modelo, se toma, de forma balanceada, la muestra dubitada del hablante contra la muestra indubitada del mismo y, también, con muestras indubitadas pertenecientes a otros hablantes.

En las figuras 4 y 5 se ilustran las instancias de entrenamiento con puntos de color rojo para la *clase 0* y azul para la *clase 1*. Igualmente, se ilustra con puntos de color amarillo las instancias de prueba de la *clase 0* y en color aguamarina la muestra dubitada comparada con otros hablantes. Se observa que el número de instancias de entrenamiento es considerablemente mayor (164,680 para el inglés y 156,646 para el húngaro) que las instancias de prueba (272 en inglés y 224 en húngaro). Ahora bien, las figuras muestran que en las instancias de entrenamiento hay un alto traslapo de los puntos, mientras que las de prueba se observa una clara diferenciación espacial con un translapo mínimo.

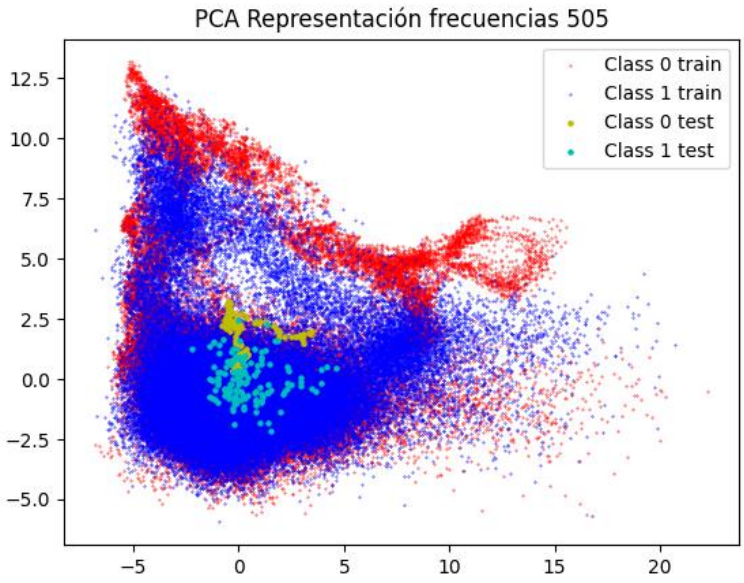


Ilustración 4. *Visualización en 2D de la representación de 505D de los datos del conjunto AusEng500 del hablante 0001 usando PCA. El color rojo representa las instancias de entrenamiento de la* clase 0*, y azul para la* clase 1. *El Color amarillo representa las instancias de prueba para la* clase 0*, y aguamarina para la* clase 1.

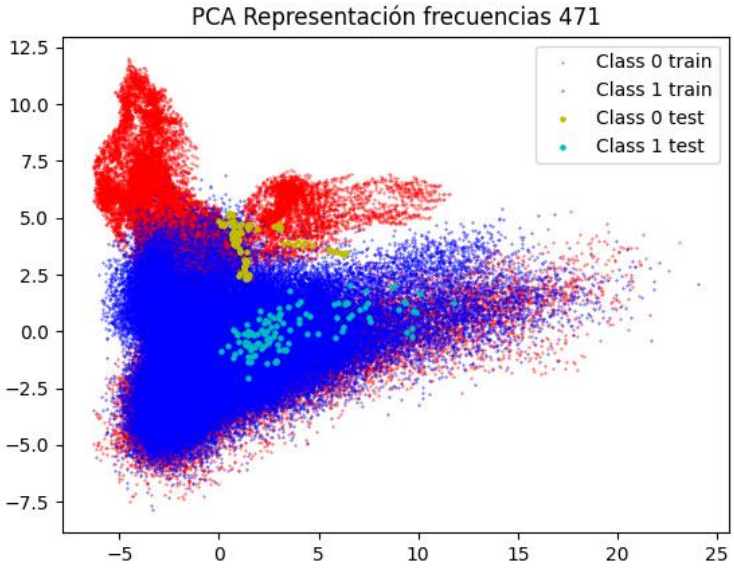


Ilustración 5. *Visualización en 2D de la representación de 471D de los datos del conjunto ForVoice120+ del hablante 001 usando PCA. El color rojo representa las instancias de entrenamiento de la* clase 0, *y azul para la* clase 1. *El Color amarillo representa las instancias de prueba para la* clase 0, *y aguamarina para la* clase 1.

6.2. Resultados del sistema de verificación de muestras dubitadas

En la Figura 6 se ilustra el histograma de los resultados del sistema de verificación usando el conjunto de parámetros definidos por defecto en la tabla 1 para los dos conjuntos de datos usados. En ambos conjuntos de datos la mayoría de los hablantes obtuvieron ratas de error de $EER < 0.05$, siendo para el inglés 122 hablantes (73%) y 101 para el húngaro (84%). El intervalo de EER más común resultó el mismo para ambas lenguas en $0.02 < EER \leq 0.03$. En el caso de el conjunto AusEng500, para esta visualización se removió un resultado atípico de $ERR = 0.332$ para el hablante No. 647.

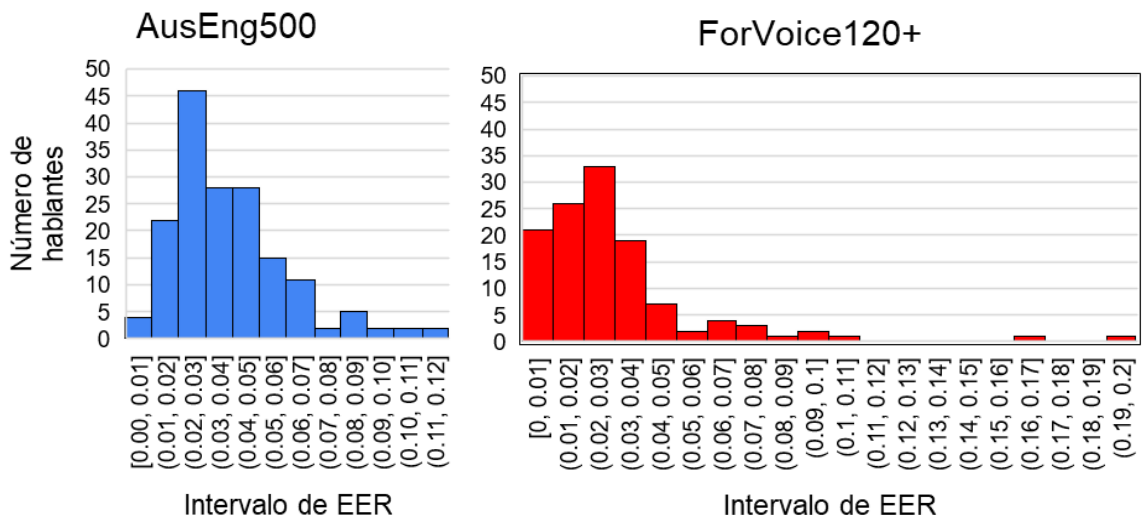


Ilustración 6. *Histograma del valor de EER por número de hablantes en el corpus AusEng500 (izquierda). Histograma del valor de EER por número de hablantes en el corpus ForVoice120+ (derecha).*

Ahora bien, para los hablantes que obtuvieron promedios de $EER \geq 0.05$ se observó que la variabilidad entre las 10 muestras dubitadas escogidas aleatoriamente es alta. Por ejemplo, en la Figura 7 se comparan los resultados de las 10 muestras extraídas de manera aleatoria del hablante 1510 en AusEng500 con un promedio de $EER = 0.106$ (izquierda) contra el hablante 2052 el cual tiene un promedio de $EER = 0.015$ (derecha). En este último se visualiza que la dispersión de los resultados es más baja.

La Figura 8 presenta los resultados de dispersión para dos hablantes de ForVoice120+. En este caso se observa como los resultados del hablante 7 quien obtuvo un $EER = 0.295$ tiene una alta dispersión mientras que el hablante 55 con un promedio de $EER = 0.011$ tiene todos los resultados cercanos a 0.

La disimilitud entre los resultados depende de las transiciones de sílabas. Cuando un hablante se expresa con enunciados cerrados y limita su oratoria, entonces el modelo no puede ser concluyente al evocar un resultado debido a que no hay características que lo alejen de la población de referencia.

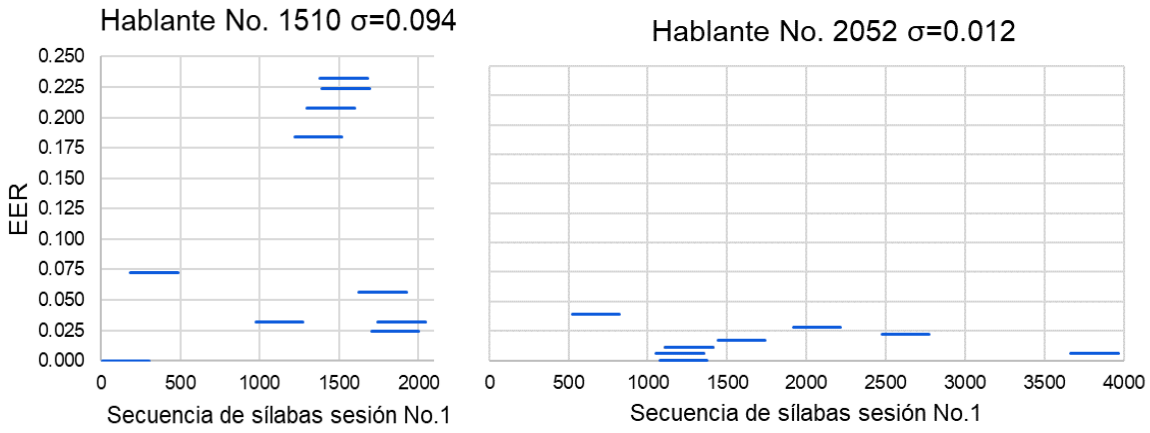


Ilustración 7. *Representación gráfica del segmento extraído (tamaño 300) de la muestra dubitada para prueba y el resultado del EER de la validación cruzada en el hablante 1510 (izquierda) y del hablante 2052 (derecha) en el corpus AusEng500.*

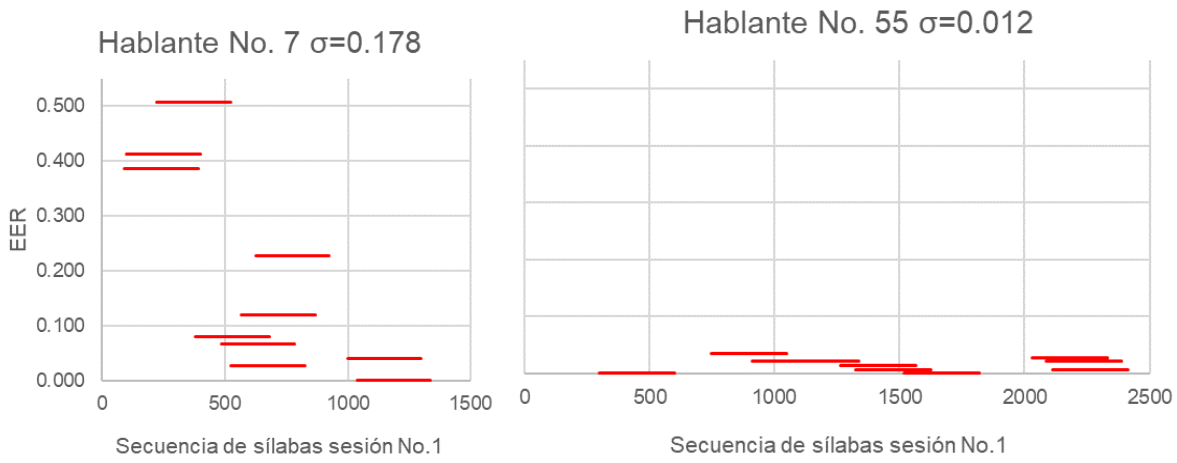


Ilustración 8. *Representación gráfica del segmento extraído (tamaño 300) de la muestra dubitada para prueba y el resultado del EER de la validación cruzada en el hablante 7 (izquierda) y del hablante 55 (derecha) en el corpus ForVoice120+.*

6.3. Análisis de parámetros

Como se explicó en el arreglo experimental, las siguientes subsecciones presentan los resultados al variar los valores de los parámetros de manera individual a partir de los valores por defecto usados.

6.3.1. Grilla de rangos frecuencias G.

La Figura 9 muestra los resultados promedio para los hablantes del sistema usando muestras seleccionadas con la semilla del generador de números aleatorios fijada en el valor 1 y los valores por defecto para los demás parámetros. El valor del tamaño de la grilla de la matriz de rangos de transiciones de frecuentas G, varía desde 10 hasta 60 mostrando un mismo patrón de tendencia para los dos conjuntos de datos. En el tope de las barras se codifica el error estándar para todos los hablantes de cada conjunto.

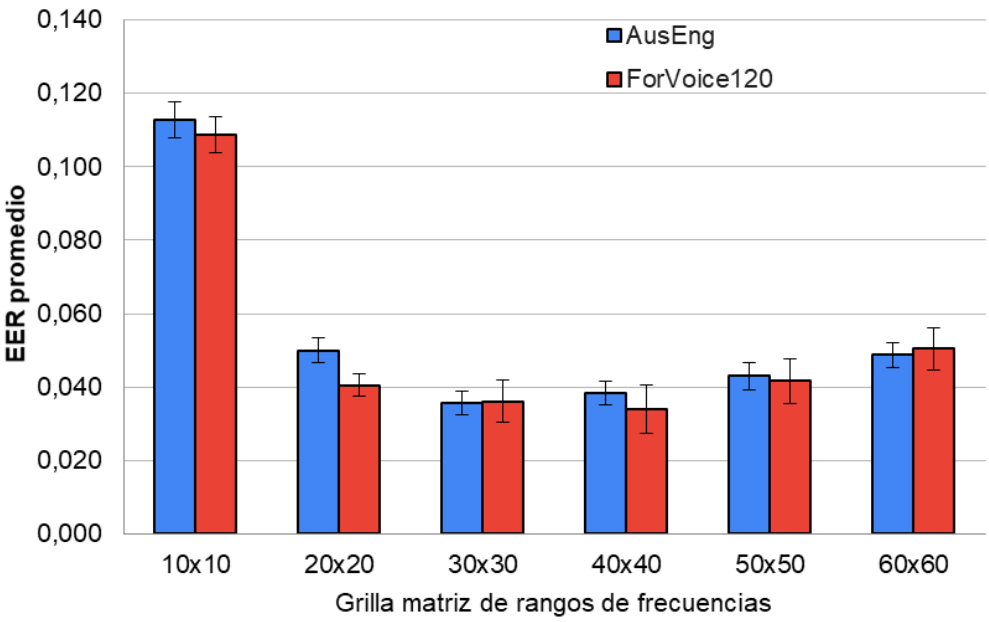


Ilustración 9. *Resultados del sistema de verificación de muestras dubitadas usando el conjunto de parámetros por defecto y variando el tamaño de la grilla de rangos de frecuencias.*

6.3.2. Tamaño de la muestras indubitadas m_i

De manera similar, en la Figura 10 se muestra una exploración de los resultados del sistema de verificación variando el tamaño de las muestras indubitadas m_i desde 300 sílabas hasta 1200.

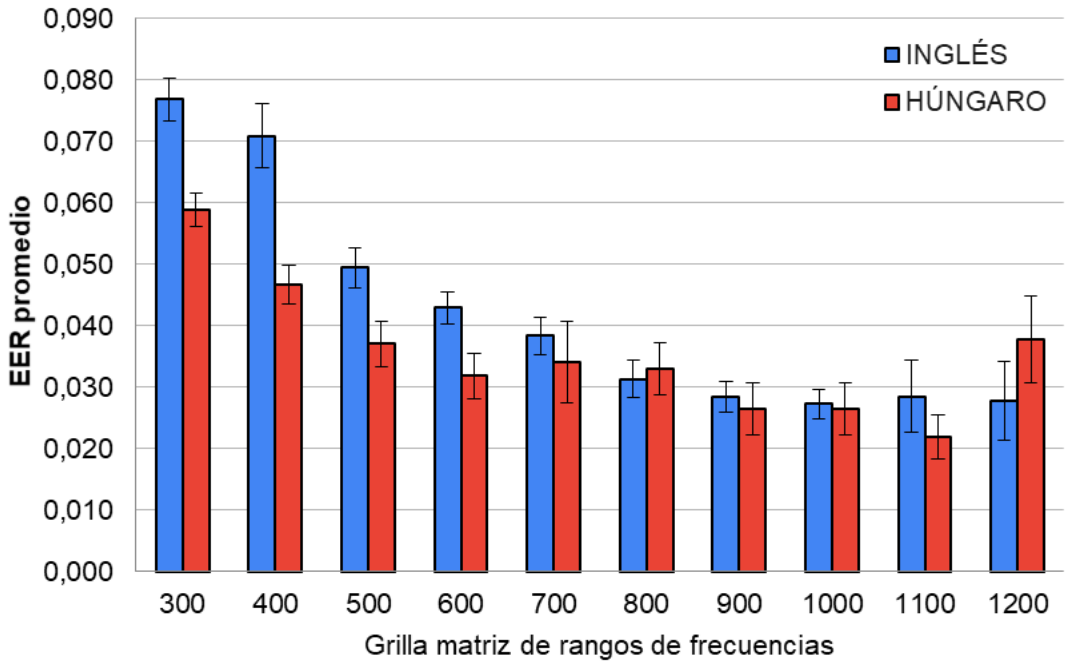


Ilustración 10. *Resultados del sistema de verificación de muestras dubitadas usando el conjunto de parámetros por defecto y variando el tamaño de las muestras indubitadas.*

6.3.3 Resultados finales.

Tomando como referencia los valores encontrados al variar los parámetros de grilla y de muestra indubitada, se obtienen los mejores resultados promedio que se visualizan en la siguiente tabla:

Tabla 2

Mejores resultados obtenidos del sistema de verificación de muestras dubitadas promediando 10 muestras tomadas aleatoriamente por hablante. Se reporta EER y estimaciones de $\widehat{\boldsymbol{C}}_{\boldsymbol{llr}}$.

Conjunto de datos	**Grilla (G)**	**Muestra indubitada (m_i)**	**EER**	$\widehat{\boldsymbol{C}}_{\boldsymbol{llr}}$
AusEng500 (inglés)	30	1,000	0.0258 (0.0020)	0.0989
ForVoice120+ (húngaro)	40	1,100	0.0275 (0.0067)	0.1074

6.4. Firmas de transiciones espectrales silábicas

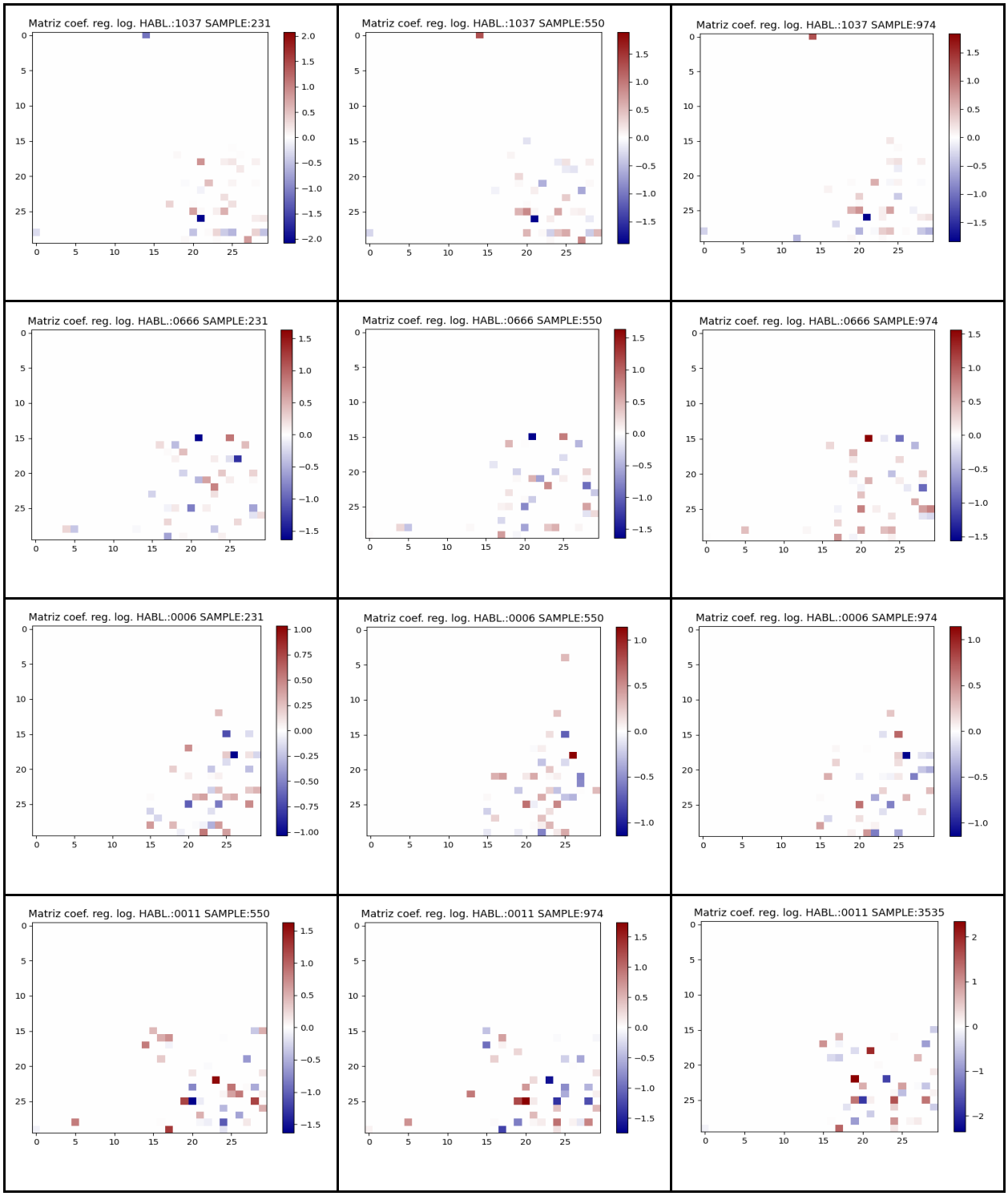


Ilustración 11. *Ejemplos de firmas de transiciones espectrales silábicas de hablantes del conjunto de datos AusEng 500 (inglés).*

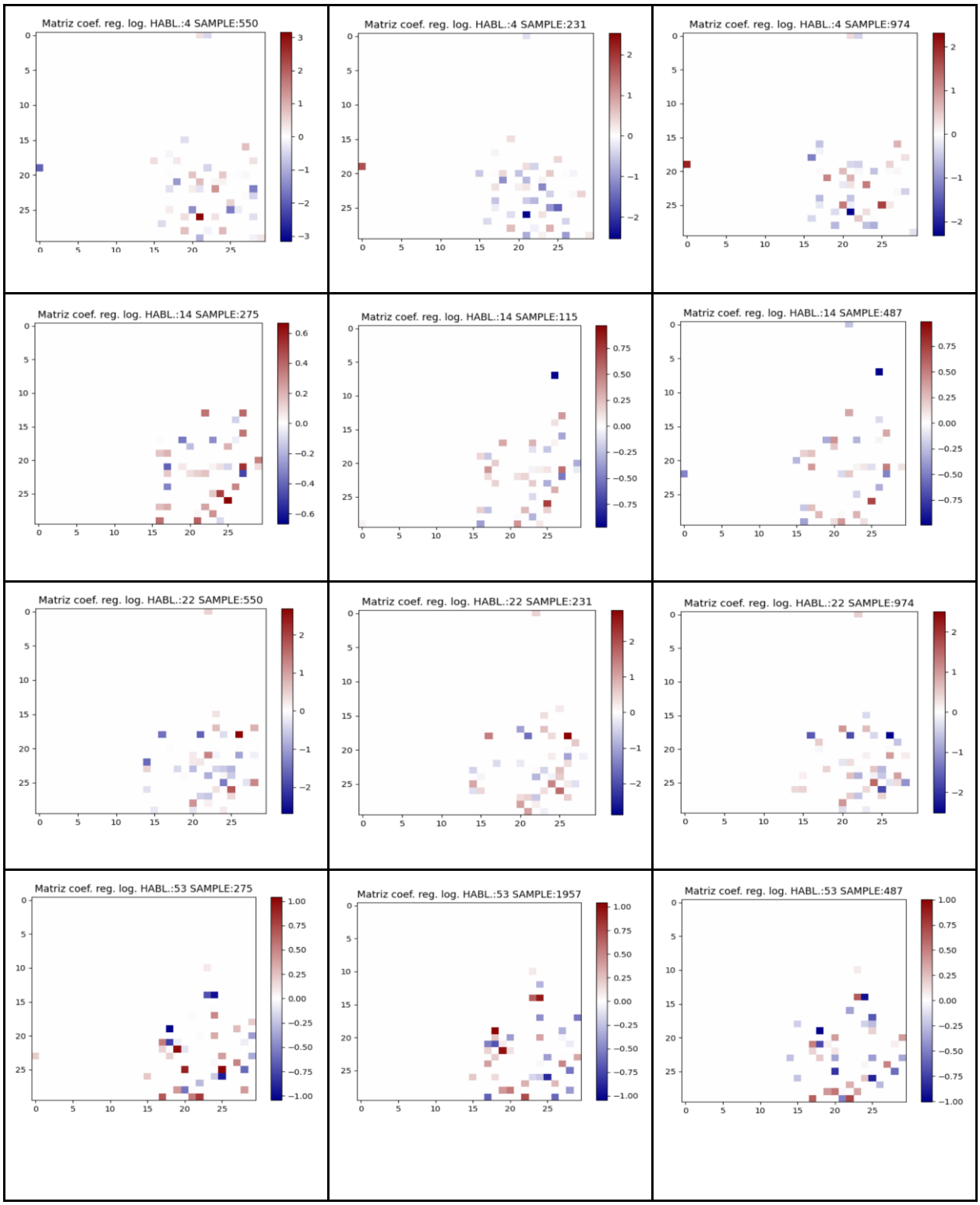


Ilustración 12. *Ejemplos de firmas de transiciones espectrales silábicas de hablantes del conjunto de datos ForVoice120+ (húngaro).*

La Figura 11 muestra, para el conjunto AusEng500, 12 firmas espectrales organizadas donde las filas muestran tres firmas correspondientes a tres diferentes muestras dubitadas para un hablante, y en se muestran resultados para 4 diferentes hablantes. Cada gráfica indica la identificación del hablante y la posición de la sílaba de donde se extrajo la muestra dubitada de 300 sílabas. Los pixeles de color rojo representan los coeficientes positivos de la regresión logística y los azules los negativos. Visualmente se aprecia que las firmas son similares en las tres gráficas de una fila (intra-hablante) y que las filas son diferenciadas entre sí (inter-hablante). La Figura 12, muestra resultados equivalentes para el conjunto de datos ForVoice120+.

7. Discusión

El aporte metodológico realizado utilizando aprendizaje de máquina con las frecuencias de las transiciones silábicas a mostrado que, efectivamente, la producción de secuencias de sílabas en el hablante es un proceso cognitivo estable que tiene potencial para la verificación de hablantes o, incluso, puede trascender a la atribución de autoría en textos escritos en aplicaciones de mensajería.

Ahora bien, el enfoque del trabajo previo (Sztahó, 2022) donde consistía en entrenar un sistema que evalúa la información semántica de las transcripciones obtuvo valores muy inferiores con respecto a los que fueron reportados con el modelo propuesto. Asimismo, la metodología aplicada donde se seleccionan 100 hablantes para entrenar el modelo para posteriormente validarlo con los 20 restantes no es adecuado porque tiene un comportamiento de identificación de emparejamiento de hablantes en grupo, siendo que en el ambiente forense el interés es la verificación individual de las muestras dubitadas e indubitadas.

El enfoque desarrollado, utiliza todos los hablantes de los conjuntos de datos para prueba para extraer muestras de las grabaciones y marcarlas como muestras dubitadas e indubitadas las cuales dependen de la clase, los resultados se reportan sobre todos los hablantes y no sobre una pequeña fracción. Con este enfoque se aprovechan al máximo los datos disponibles para verificar si un par de muestra dubitada e indubitada tienen un mismo origen. El algoritmo de predicción fue llevado al extremo al utilizar el esquema de evaluación *leave-one-sample-out*.

7.1 Representación por PCA

Sobre los resultados obtenidos con el análisis usando PCA (figuras 4 y 5) se puede apreciar que, a pesar de que existe un alto traslapo de los pares de muestras de *clase 0* y *clase 1* en el conjunto de entrenamiento (puntos azules y rojos), hay un patrón diferencial de los pares de muestras de diferente hablante (*clase 1* en azul) los cuales se agrupan alrededor del punto de acumulación de la mayoría de pares de muestras (origen). Por otro lado, los pares de muestras correspondientes al mismo hablante (*clase 0* en rojo) se ubican en la periferia. Con respecto a los conjuntos de prueba, se observa que aunque también se agrupan cerca del punto de origen, si hay una distinción entre ambas clases. Es entonces se puede concluir que la representación vectorial a partir de transiciones de rangos de frecuencias de sílabas permite diferenciar los pares de muestras de un mismo hablante de los de diferentes hablantes para una muestra dubitada particular.

7.2 Resultados de rendimiento en la verificación de muestras

La mayoría de los valores de EER obtenidos en la Figura 6 muestran que están por debajo del 5% lo que resulta comparable al valor crítico p-value (valor que en estadística

cuantifica el grado de duda). Sin embargo, no hay un criterio en lo referente a la comparación de hablantes donde se establezca un valor mínimo de confianza. Así, se toma “prestado” el concepto de la estadística y la ciencia donde se indica que valores menores de 0.05 otorgan un alto nivel de confianza. Establecido ese criterio, observamos que el sistema de verificación logra una comparación fiable de las muestras otorgando resultados que diferencian al hablante en cuestión de los demás presentados en el conjunto de datos.

Los resultados obtenidos desde el nuevo enfoque muestran una mejora significativa al estudio realizado por Sztahó (2022), esto se debe en gran medida a que se utilizan las frecuencias y transiciones de sílabas en lugar de utilizar la clasificación semántica de las muestras. De esta manera, se corrobora la teoría de que la selección de palabras cumple un papel importante al momento de perfilar a un hablante y que su uso léxico está condicionado por las experiencias y el vocabulario que la persona tenga a nivel cognitivo. También se observa que el modelo es independiente de la lengua y no es necesario etiquetar las palabras de estudio con herramientas dependiente de la lengua como etiquetadores gramaticales. En nuestro caso, solamente se requiere un corpus de referencia de la población para lograr extrapolar las sílabas en valores numéricos que son cuantificables para la conformación de los vectores matriciales.

Sobre las figuras 7 y 8, se observa que las muestras extraídas dependen del hablante, pero existen ciertos casos los cuales sin importar desde donde se extrae la muestra dubitada que es utilizada únicamente para prueba, los resultados de confiabilidad serán buenos. Según los valores de EER obtenidos, para la mayoría de los hablantes (>70%) cualquier muestra tomada aleatoriamente de la sesión 1 es útil para el proceso de verificación.

En general, se puede concluir que con una muestra dubitada de tamaño 300, el modelo propuesto puede verificar exitosamente a un 70% de la población. Si esta verificación no es posible, se requiere tener material dubitado de mayor extensión (ej, 2,000 sílabas) y de este extraer secuencialmente muestras de 300 sílabas y que, teóricamente, con un 100% de certeza se encontrará una muestra dubitada verificable.

7.3 Resultados con la variación de parámetros

La Figura 9 muestra los resultados de la exploración del parámetro G. Es claro que, cuando se utilizan valores muy pequeños, los valores de EER son muy altos, mostrando que, al reducir mucho el número de intervalos de frecuencias de sílabas, la escala se vuelve muy "gruesa" y hay pérdida de información, la cual se ve reflejada en el bajo rendimiento del sistema. Por el contrario, cuando se aumenta G a valores por encima de 50, dado que el tamaño de la matriz de transiciones de frecuencias aumenta de manera cuadrática, la representación de los datos se vuelve muy dispersa perjudicando así el proceso de clasificación. El valor adecuado para el tamaño de la grilla de clasificación está entre 30 y 40.

En la Figura 10 se observa que cuando se segmentan las muestras indubitadas en tamaños pequeños tiene un resultado negativo para el rendimiento del sistema al momento de comparar el hablante. Esto ocurre porque entre conjuntos más pequeños cercanos al valor del traslapo, existe un sobre-entrenamiento del algoritmo y se vuelve más riguroso para poder otorgar un valor de identificación y puede incurrir en falsos negativos. Por otro lado, cuando se usan agrupaciones muy grandes se reduce el número de muestras indubitadas posibles a obtener de las muestras de laboratorio del indicado (sesión 2). Esto reduce el número de pares de muestras que se pueden formar para los datos de entrenamiento y de prueba, lo que puede llegar a perjudicar el rendimiento en la clasificación. Además, esto también limita el número

de muestras indubitadas que se pueden obtener de los datos de los hablantes de contraste, reduciendo aún más el número de pares de muestras que se pueden obtener. En conclusión, el mejor valor para el agrupamiento de la muestra indubitada es alrededor de 1000.

7.4. Interpretabilidad de las firmas espectrales

Los resultados de las figuras 11 y 12 muestran que visualmente es posible diferenciar las firmas espectrales entre hablantes y que existe una alta similitud visual entre las diferentes firmas de un hablante. Siendo los ejes de las firmas completamente interpretables, dado que es una escala logarítmica de frecuencias de sílabas tomada de un corpus de contraste (i.e. Twitter), la interpretabilidad de las firmas es directa. Además, cada "*pixel*" de la firma representa una transición de la frecuencia codificada en la horizontal a la frecuencia en la vertical. La intensidad del color representa la importancia de esta transición para el hablante. Dada la simplicidad de la interpretación estas firmas tienen el potencial de ser admisibles en ambientes forenses como material probatorio verificable por personal no técnico y eventualmente por abogados y jueces.

7.5. Comparación con otros estudios

Debido a que en la literatura solamente se encontró como línea de base el trabajo desarrollado por Sztahó (2022), se realiza la comparación del EER contra los resultados arrojados cuando utiliza los conjuntos de ForVoice120 y AusEng500 sin hacer división de los párrafos, el resultado es arrojado en las siguientes tablas:

Tabla 3

Comparación con los mejores resultados de otros estudios para el conjunto de datos AusEng500 (inglés australiano).

Sistema o estudio	Descripción	EER	EER este estudio	Diferencia EER
Weber et al., (2022)	Audios usando programa Phonexia XL3	0.022	0.0258	-0.0038
Sztahó, D., & Fejes, A. (2023)	Audios y deeplearning usando las técnicas *speaker enrollment* ECAPA-TDNN	0.016	0.0258	-0.0098
Sigona & Grimaldi, (2023)	Audios usando SYS4 ECAPA-TDNN	0.020	0.0258	-0.0058

Tabla 4

Comparación con los mejores resultados de otros estudios para el conjunto de datos ForVoice 120+ (húngaro).

Sistema o estudio	Descripción	EER	EER este estudio	Diferencia EER
Sztahó et al. (2022)	Usando transcripciones y representación semántica con Doc2Vec	0.34	0.0275	0.3125
Sztahó, D., & Fejes, A. (2023)	Audios y deeplearning *speaker enrolment* ECAPA-TDNN	0.010	0.0275	-0.0175
Abed, M. H., & Sztahó, D. , (2023)	ECAPA-TDNN alineando emociones	0.026	0.0275	-0.0015

Como se puede observar los resultados arrojados utilizando la frecuencia léxica son mejores que los que se obtuvieron utilizando Doc2Vec. Sin embargo, Sztahó (2023) utilizó en un estudio posterior los audios. La comparación presentada en las tablas 2 y tres muestra que el método presentado en este estudio, mejoró considerablemente el único trabajo anterior que abordó la tarea de identificación de hablantes usando transcripciones de los audios (Sztahó et

al.,2022) EER=0.34. Los bajos resultados de este estudio prácticamente descartaban el uso de transcripciones como información para la identificación forense. Nuestros resultados con EER=0.0275 refutan estos resultados mostrando que usando las transcripciones de los audios se puede obtener un rendimiento comparable a los sistemas basados en audio. Consideramos que esta amplia diferencia se debe a que Sztahó et al. (2022) usaron una representación semántica la cual resulta inconveniente para identificación, mientras que la representación usada en este estudio tiene motivaciones cognitivo-lingüísticas las cuales mostraron su eficacia en la tarea de verificación forense.

Adicionalmente, al comparar el rendimiento en la tarea de identificación forense usando audios versus el uso de transcripciones, consideramos que aunque el uso de audios produce menores ratas de EER que el uso de transcripciones, esa diferencia es marginal. Además, se debe poner en contexto que el estado del arte con el que nos comparamos está basado en el método ECAPA-TDNN el cual se basa en aprendizaje profundo (deeplearning), implicando que los modelos tienen un gran número de parámetros del orden de decenas de millones (Weber et al., 2022). En contraste el modelo presentado basado en la representación de transiciones espectrales silábicas y regresión logística el número de parámetros es significativamente menor, esto es 346 para AusEng500 y 471 para ForVoice120+. Siendo así, las diferencias en EER menores de 0.018 no son considerables teniendo en cuenta las ventajas de simplicidad e interpretabilidad de las transiciones espectrales silábicas.

Consideramos que los resultados muestran que la hipótesis subyacente de la representación de transiciones espectrales silábicas de que las transiciones entre diferentes zonas del inventario silábico de un hablante son individualizantes. Esto implicaría que, dado que en el habla espontánea los hablantes producen una secuencia de sílabas de manera

eficiente, a nivel cognitivo las diferentes zonas del inventario silábico deberían estar interconectadas para lograr dicha eficiencia. Este estudio provee evidencia indirecta de que dicha interconexión es diferente para cada individuo.

8. Conclusiones

Los resultados demuestran la factibilidad del uso de modelos computacionales simples para la indexación de las sílabas extraídas a partir de la transcripción de audios que simulan el ambiente forense. La contribución más significativa es la interpretabilidad del modelo ya que utiliza la regresión logística en combinación con la nueva representación de *transiciones de rangos de frecuencias silábicas* como modelo predictivo que busca aquellas características importantes generadas por un locutor y que comparte o lo diferencia de los locutores que se evalúa desde una perspectiva psicolingüística.

Los bajos registros de EER en las pruebas demuestran que puede ser utilizado como herramienta automática complementaria a los sistemas actuales de reconocimiento de hablantes. Ahora bien, el desarrollo del presente proyecto demuestra que hay una relación intrínseca entre la selección de palabras y el perfilamiento de los locutores. Se logra observar, por tanto, que incluso aunque las personas tienen un conocimiento transversal de la lengua, su selección es distintiva logrando conseguir una perfilación por la frecuencia de las palabras en un corpus significativo, la cual se ve reflejada en la secuencia de sílabas que producen.

El modelo proporcionado en la presente investigación demuestra que el algoritmo funciona de manera independiente de la lengua toda vez que solamente es necesario crear las matrices con un corpus de referencia para ser el medio de comparación de las frecuencias silábicas.

9. Trabajos a futuro

El sistema muestra el potencial discriminatorio de las secuencias y frecuencias silábicas para otras tareas del Procesamiento del Lenguaje Natural y la Lingüística Computacional. Este método puede utilizarse incluso en textos escritos para la verificación de autoría, logrando abarcar áreas más grandes de la lingüística forense, permitiendo estudios más robustos, automatizados e interpretables.

Potencial de la representación para otros tipos de identificación de género, edad, dialecto a partir de muestras de habla.

De este trabajo se derivan nuevas perspectivas de investigación como el indagar cuales son las razones y características del por qué algunas muestras dubitadas son más convenientes que otras para la identificación.

Referencias

Abed, M. H., & Sztahó, D. (2023). Effects of emotional speech on forensic voice comparison using deep speaker embeddings. Magyar Számítógépes Nyelvészeti Konferencia, 19, 159-170.

Arons, B. (1992). A review of the cocktail party effect. Journal of the American Voice I/O society, 12(7), 35-50.

Beke. (2021). ForVOICE 120+: magyar nyelvutánkovetéses adatbázis kriminalisztikai célú hangosszehasonlításra.

Bhattacharyya, D., Ranjan, R., Alisherov, F., Choi, M., et al. (2009). Biometric authentication: A review. International Journal of u-and e-Service, Science and Technology.

Bloch, Bernard. (1948). A set of postulates for phonemic analysis. Language, 24(1), 3-46.

Brendel, B., Erb, M., Riecker, A., Grodd, W., Ackermann, H., & Ziegler, W. (2011). Do we have a “mental syllabary” in the brain? An fMRI study. Motor Control, 15(1), 34-51.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. Current Directions in Psychological Science.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. Experimental psychology.

Casas Gómez, M. (2008). El concepto de significante en el funcionalismo semántico. Romanische Forschungen, 120(3), 283-306.

Casas Gómez, M. (2008). El concepto de significante en el funcionalismo semántico. Romanische Forschungen, 120(3), 283-306.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: The Mit Press

Chomsky, N. (1957). "Syntactic Structures." Mouton.

Cholin, J., Levelt, W. J., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, *99*(2), 205-235.

Coulthard, M. (2016). An introduction to forensic linguistics: language in evidence. Routledge.

Eriksson, A. (2012). Aural/acoustic vs. automatic methods in forensic phonetic case work. Forensic Speaker Recognition: Law Enforcement and Counter-terrorism, pages 41–69.

Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., & Smith, N. (2015). Sparse overcomplete word vector representations. arXiv preprint arXiv:1506.02004.

Garcia, A. M., & Martin, J. C. (2006). Function words in authorship attribution studies. Literary and Linguistic Computing, 22(1), 49-66.

Garrett, B. L., & Rudin, C. (2023). Interpretable algorithmic forensics. Proceedings of the National Academy of Sciences, 120(41), e2301842120.

Gimenes, M., & New, B. (2015). Worldlex: Twitter and blog word frequencies for 66 languages. Behavior research methods, 1-10.

Gong, Y., Khurana, S., Karlinsky, L., & Glass, J. (2023). Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. arXiv preprint arXiv:2307.03183.

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T.E., Beke, A., Neuberger, T., & Nikléczy, P. (2012). BEA: Beszélt nyelvi adatbázis.

Graves WW, Grabowski TJ, Mehta S, Gordon JK (2007): A neural signature of phonological access: Distinguishing the effects of word frequency from familiarity and length in overt picture naming. J Cogn Neurosci 19:617–631

Hernández Pina, F. (1980). Las relaciones entre pensamiento según Piaget, Vygotsky, Luria y Bruner. In Anales de la Universidad de Murcia. Filosofía y Letras. Murcia: Universidad, Secretariado de Publicaciones.

Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. Information Fusion, 94, 17-31.

Jimenez, S., Avila, Y., Dueñas, G., & Gelbukh, A. (2020). Automatic prediction of citability of scientific articles by stylometry of their titles and abstracts. Scientometrics, 125(3):3187–3232.

Labov, W. (1972). Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press

Lan, M., Sung, S. Y., Low, H. B., & Tan, C. L. (2005, July). A comparative study on term weighting schemes for text categorization. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. (Vol. 1, pp. 546-551). IEEE.

LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.

Levelt, W. J., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary?. Cognition, 50(1-3), 239-269.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1–75.

Le, Q.,& Mikolov, T. (2014). Distributed representations of sentences and documents. 31st International Conference on Machine Learning, ICML 2014, 4.

Maltoni, D. (2003). Handbook of Fingerprint Recognition. Springer professional computing. Springer.

Martinc, M., Skrjanec, I., Zupan, K., & Pollak, S. (2017). PAN 2017: Author Profiling-Gender and Language Variety Prediction. In CLEF (working notes).

Morrison G.S., Rose P, & Zhang C. (2012). Protocol for the collection of databases of recordings for forensic-voice- comparison research and practice. Aust J Forensic Sci. 44(2):155–67.

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. Science & Justice, 54(3), 245-256.

Morrison G.S., Zhang C., Enzinger E., Ochoa F., Bleach D., Johnson M., Folkes B. K., De Souza S., Cummins N., Chow D., Szczekulska A. (2021). Forensic database of voice recordings of 500+ Australian English speakers (AusEng 500+)

Morrison, G.S., Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz W.F., Assmann P.F. (Eds.) The Routledge Handbook of Phonetics(ch. 21, pp. 599–634). Abingdon, UK: Taylor & Francis. https://doi.org/10.4324/9780429056253-22

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., ... & Anonymous, B. (2021a). Consensus on validation of forensic voice comparison. Science & Justice, 61(3), 299-309.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. ACM Computing Surveys (CSuR), 50(6):1–36.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR.

Romero, C. D. (2001). La identificación de locutores en el ámbito forense (Doctoral dissertation, Universidad Complutense de Madrid).

Rose, P. & Morrison, G. (2009). A response to the uk position statement on forensic speaker comparison. International Journal of Speech Language and The Law - INT J SPEECH LANG LAW, 16.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

San Segundo E., J. A. Mompeán (2017). A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity. Journal of Voice 31(5). 644-e11.

San Segundo, E.; P. Foulkes, P. French, P. Harrison, V. Hughes & Y C. Kavanagh (2018). The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. Journal of the International Phonetic Association, pp. 1-28.

San Segundo, E.; Univaso, P.; Gurlekian, J.(2019). SISTEMA MULTIPARAMÉTRICO PARA LA COMPARACIÓN FORENSE DE HABLANTES. Estudios de Fonética Experimental, ISSN 1575-5533, XXVIII, PP. 13-45

Saussure, F. (1990) Curso de lingüística general. Madrid: Alianza

Saussure, F. D. (1978). Curso de lingüística general. In Curso de lingüística general, pp. 378-378.

Schiller, N.O. (1997). The role of the syllable in speech production. Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. PhD dissertation, Nijmegen University (MPI series; 2).

Schiller, N. O. (2021). The Mental Lexicon. Oxford University Press.

Sigona, F., & Grimaldi, M. (2023). Validation of an ECAPA-TDNN system for Forensic Automatic Speaker Recognition under case work conditions. arXiv preprint arXiv:2305.10805.

Simpson, A. P. (2013). Spontaneous speech. The Bloomsbury companion to phonetics, 155-169.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, pp. 538−566.

Stefanova Spassova, M. (2009). El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español. Universitat Pompeu Fabra.

Szekrényes, I. (2014). Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. Journal on Multimodal User Interfaces 8, pages 143–150.

Sztahó, D., Beke, A., Szaszák, G., & Fejes, A. (2022). Forensic Authorship Classification by Paragraph Vectors of Speech Transcriptions. Berend Gábor, Gosztolya Gábor és Vincze Veronika (szerk.). XVII. Magyar Számítógépes Nyelvészeti Konferencia, 275-288.

Sztahó, D., & Fejes, A. (2023). Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. Journal of Forensic Sciences, 68(3), 871-883.

Velásquez, F., Godoy, J., Falcón, M., De Paz, J., Chávez, A., & Sierra, J. (2020). Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático. Res. Comput. Sci., 149(11), 91-101.

Weber, P., Enzinger, E., Labrador, B., Lozano-Díez, A., Ramos, D., González-Rodríguez, J., & Morrison, G. S. (2022). Validations of an alpha version of the E3 Forensic Speech Science System (E3FS3) core software tools. Forensic Science International: Synergy, 4, 100223.

Wilson, S. M., Isenberg, A. L., & Hickok, G. (2009). Neural correlates of word production stages delineated by parametric modulation of psycholinguistic variables. Human brain mapping, 30(11), 3596-3608.