Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada,
Vol. 23, n.º 2, julio-diciembre 2024, pp. 1069-1097,
ISSN-L 1729-9721; eISSN: 2413-2659
https://doi.org/10.15381/lengsoc.v23i2.29209

Un estudio de caso sobre los cuatro momentos espectrales y el pico de máxima intensidad de /s/ en una voz natural y una voz clonada por la IA Elevenlabs[1]

A case study on the four spectral moments and the peak intensity of /s/ in a natural voice and a voice cloned by AI Elevenlabs

Um estudo de caso sobre os quatro momentos espectrais e o pico máximo de intensidade de /s/ em uma voz natural e uma voz clonada pela AI Elevenlabs

Fernando Aarón Torres Castillo
Universidad Nacional Mayor de San Marcos, Lima, Perú
fernando.torres2@unmsm.edu.pe
https://orcid.org/0000-0002-1432-8811

Oscar Esaul Cueva Sanchez
Universidad Nacional Mayor de San Marcos, Lima, Perú
oscar.cueva1@unmsm.edu.pe
https://orcid.org/0000-0003-1361-2367

Jhon Jimenez Peña
Universidad Nacional Mayor de San Marcos, Lima, Perú
jhon.jimenez@unmsm.edu.pe
https://orcid.org/0000-0003-3317-6152

Erika Amalec Shicshi Romero
Universidad Nacional Mayor de San Marcos, Lima, Perú
erika.shicshi@unmsm.edu.pe
https://orcid.org/0000-0002-0431-9591

Resumen

La presente investigación compara los cuatro momentos espectrales y el pico de mayor intensidad de la fricativa alveolar [s] en posición de coda, entre la voz natural y la artificial de un locutor. Los datos de la voz natural fueron recolectados en un entorno controlado y segmentados posteriormente con el *software* Praat. Para la voz artificial, se utilizó la tecnología de clonación de voz de ElevenLabs. El análisis de varianza muestra diferencias significativas entre las medias de los parámetros acústicos analizados —centro de gravedad, desviación estándar, curtosis, asimetría y el pico de máxima intensidad— en la voz natural y artificial. Sin embargo, al analizar los datos agrupados por la vocal que precede a la fricativa, se observa que no en todos los contextos existen diferencias significativas. Los resultados indican que los parámetros espectrales de la fricativa alveolar /s/ —especialmente el centro de gravedad— son útiles para distinguir entre la voz natural y su contraparte artificial.

Palabras clave: lingüística forense; fonética forense; Praat; Anova; estadística.

1 Esta investigación fue impulsada por el Gabinete de Lingüística Forense —del Instituto de Investigación de Lingüística Aplicada (CILA) de la Universidad Nacional Mayor de San Marcos—, el cual fue creado mediante la Resolución Decanal n.° 000623-2021-D-FLCH/UNMSM, también por el *Voice Deepfakes Research Project*, el cual es financiado por MCIN/AEI/10.13039/501100011033 y por FEDER.

Abstract

This research compares the four spectral moments and the peak of greatest intensity of the alveolar fricative [s] in coda position between the natural and artificial voice of a speaker. The data from the natural voice were collected in a controlled environment and subsequently segmented with Praat software. For the artificial voice, ElevenLabs' voice cloning technology was used. The analysis of variance shows significant differences between the means of the acoustic parameters analyzed—center of gravity, standard deviation, kurtosis, skewness, and the peak of maximum intensity—in the natural and artificial voices. However, when analyzing the data grouped by the vowel preceding the fricative, it is observed that not all contexts show significant differences. The results indicate that the spectral parameters of the alveolar fricative /s/—especially the center of gravity—are useful to distinguish between the natural voice and its artificial counterpart.

Keywords: forensic linguistics; forensic phonetics; Praat; Anova; statistics.

Resumo

A presente investigação compara os quatro momentos espectrais e o pico de maior intensidade da fricativa alveolar [s] em posição de coda, entre a voz natural e artificial de um falante. Os dados de voz natural foram coletados em ambiente controlado e posteriormente segmentados com software Praat. Para a voz artificial foi utilizada a tecnologia de clonagem de voz da ElevenLabs. A análise de variância mostra diferenças significativas entre as médias dos parâmetros acústicos analisados – centro de gravidade, desvio padrão, curtose, assimetria e pico de intensidade máxima – na voz natural e artificial. Porém, ao analisar os dados agrupados pela vogal que antecede a fricativa, observa-se que nem em todos os contextos existem diferenças significativas. Os resultados indicam que os parâmetros espectrais da fricativa alveolar /s/ – especialmente o centro de gravidade – são úteis na distinção entre a fala natural e a sua contraparte artificial.

Palavras-chave: linguística forense; fonética forense; Praat; Anova; estatística.

Recibido: 10/02/2024 Aceptado: 26/07/2024 Publicado: 30/12/2024

1. Introducción

El *deepfake* es una técnica que utiliza algoritmos de inteligencia artificial y redes neuronales de aprendizaje profundo para crear archivos artificiales de vídeo, audio e imagen a partir de rostros y voces reales. Si bien es cierto, actualmente, la técnica descrita se usa para fines beneficiosos para la sociedad (académicos, culturales, de entretenimiento, etc.), también se utiliza para cometer delitos. Un mal uso del *deepfake* es la clonación de voz, empleada para suplantar la identidad de una persona y cometer crímenes. A nivel mundial, se han reportado diversos delitos de este tipo como el caso reciente del presidente Joe biden en Estados Unidos de Norteamérica[2]. En el audio, supuestamente, se escucha al mandatario pedir a los ciudadanos de New Hampshire que no vayan a votar en las elecciones primarias que se realizaron el 23 de enero de 2024.

En el Perú, la Autoridad Nacional de Protección de Datos Personales (ANPD) del Ministerio de Justicia y Derechos Humanos (MINJUSDH) sostiene que la ciberdelincuencia y la suplantación de identidad se ha incrementado. Según esta entidad, se han registrado casos de estafas y de otros delitos donde se implementaron sistemas de inteligencia artificial (IA) para clonar la voz[3]. Ahora bien, la clonación de voz es un proceso que está al alcance de cualquier persona. Jimenez *et al.* (2024) advierten que la voz artificial de personajes conocidos puede ser clonada con total facilidad a partir de programas de libre acceso o paga, como lo es FakeYou, un convertidor de texto a voz. Además, San Segundo (2023) señala que las «clonaciones artificiales de la voz de una persona [son] tan realistas que se han convertido en los nuevos "mejores impostores" de una voz real. Los deepfakes son voces extremadamente similares a las humanas» (p. 129).

2 https://elpais.com/internacional/elecciones-usa/2024-01-23/biden-suplantado-con-inteligencia-artificial-para-interferir-en-las-primarias-de-new-hampshire.html

3 https://www.gob.pe/institucion/minjus/noticias/805436-autoridad-nacional-de-proteccion-de-datos-personales-brinda-recomendaciones-ante-nueva-modalidad-de-estafa-por-clonacion-de-voz

Ante el avance tecnológico de los *deepfakes* y el incremento de delitos por el uso de estos, se necesita «implementar una metodología que identifique cómo distinguir qué muestras de voz son deepfakes y cuáles son reales» (San Segundo y Gibson, 2024, p. 1). Los estudios sobre las diferencias entre voces artificiales y reales son limitados. Para lograr obtener tal metodología, es necesario identificar qué características fonético-acústicas permitirán distinguir adecuadamente una voz artificial de una voz natural. Como señalan San Segundo y Delgado (2024), existen estudios de caso que analizan una voz natural y su versión clonada, así como investigaciones que identifican características específicas que permiten diferenciar entre voces reales y clonadas.

Esta investigación se inscribe entre estos estudios, puesto que busca analizar la variación del segmento fricativo alveolar sordo /s/ en posición de coda en una voz natural y compararla con la variación de su contraparte artificial, con el fin de evaluar su capacidad discriminativa entre ambas voces. En ese sentido, el objetivo es examinar cinco parámetros acústicos del sonido [s] —los cuatro primeros momentos espectrales y el pico de mayor intensidad— para identificar cuáles de ellos pueden utilizarse en la comparación forense de voz.

2. Marco conceptual

2.1. *Deepfakes* de voz y fonética forense

San Segundo y Delgado (2024) señalan que «un *deepfake* es una voz sintética (generada a partir de modelos de aprendizaje profundo) que presenta un parecido extremo con una voz real y, por tanto, se puede usar para clonar voces y suplantar la identidad de un hablante» (p. 2).

El análisis de los *deepfakes* de voz recae en una disciplina lingüística específica: la fonética forense. Esta área de estudio tiene como objetivo principal comparar voces para determinar si corresponden a una misma persona. En una definición amplia, la fonética forense se encarga de la aplicación de conceptos y métodos de la fonética en general a la investigación y resolución de delitos en los que está implicada la voz; en una definición más estrecha, la fonética forense se refiere al uso de una voz como prueba en un contexto judicial (San Segundo, 2023). En ese sentido, una de las tareas principales de la fonética forense es «discernir con el mayor grado de fiabilidad posible si concurren suficientes indicios como para sostener que dos voces pueden corresponder a la misma persona o si, por el contrario, hay que rechazar esta posibilidad» (Fernández, 2007, p. 49). La comparación de voces se logra a través del análisis acústico minucioso y exhaustivo de las voces presentes en la muestra dubitada —de la cual se desconoce a quién corresponde— y la muestra indubitada —de la cual se conoce su autoría— (Fernández, 2007). En el presente estudio —que prevé un posible caso de clonación de voz—, la muestra indubitada sería la voz natural y la muestra dubitada, la voz artificial.

Por otro lado, los estudios sobre las *deepfakes* de voz aún son escasos, principalmente en el contexto nacional. Por ejemplo, en Jimenez *et al.* (2024) se analiza las similitudes y diferencias fonéticas entre una voz natural y una voz artificial, con el objetivo de apoyar casos judiciales de clonación de voz por IA y resaltar el valor de la lingüística en el sistema judicial. Comparando la voz del narrador Mariano Closs y su versión clonada en FakeYou, los programas automáticos detectaron alta similitud, pero el análisis fonético mostró diferencias en la entonación y en la producción de ciertos sonidos —como la vibrante múltiple—, indicando que la voz artificial aún no es idéntica a la natural en términos fonéticos.

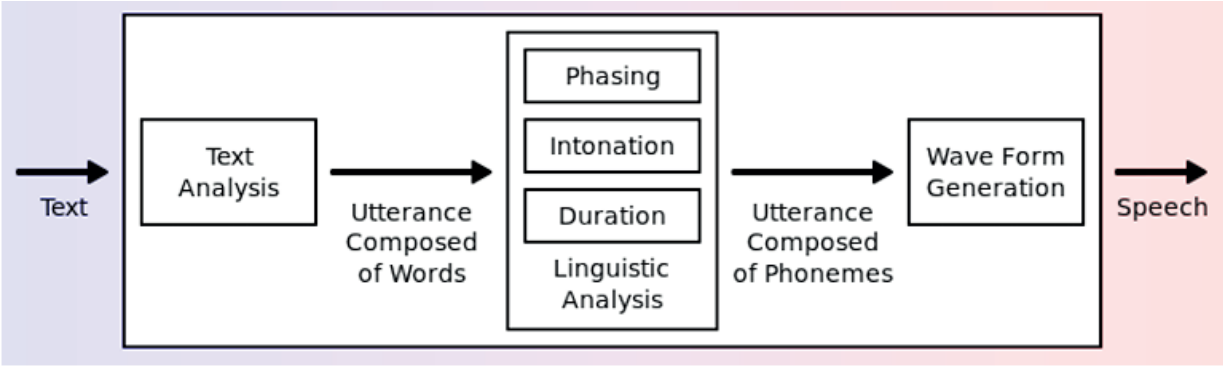
En cuanto a estudios internacionales, en San Segundo y Delgado (2024) se analiza las voces de dos pares de gemelos idénticos, seleccionados por edad y similitud en distancia euclidiana. El estudio utilizó 12 muestras de voz (2 naturales y 1 clonada por IA para cada gemelo), evaluando 21 parámetros acústicos con el *software* Praat. Los resultados indican que existe variabilidad tanto entre los gemelos (intragemelar) como dentro de cada sujeto (intrasujeto) en la mayoría de los parámetros de las voces reales. En ese sentido, la variabilidad es siempre mayor en las voces reales que en las voces clonadas, especialmente en los parámetros relacionados con la frecuencia fundamental (F0), mientras que esta diferencia es menor en los parámetros de perturbación de amplitud.

Es así como «a partir del análisis del habla, la fonética forense encuentra patrones particulares y recurrentes que responden al idiolecto de un individuo y que contribuyen a su diferenciación respecto a otras voces» (Jimenez *et al*., 2022, p. 437). Ambos estudios sobre *deepfakes* de voz señalan que la ausencia de variación en las voces clonadas podría ser un elemento relevante en la fonética forense para distinguir voces humanas, debido a que la variabilidad es una característica intrínseca de las voces naturales.

2.2. La conversión de texto en habla

Peñas (2023) describe la síntesis del habla como la tecnología que produce de manera artificial el habla humana. Así, a través de esta técnica, es posible ingresar un texto y obtener como resultado una voz artificialmente sintetizada. Asimismo, según Llisterri *et al*. (2004), la conversión de texto a habla permite que un ordenador transforme cualquier texto escrito en discurso oral, utilizando módulos especializados que procesan datos de naturaleza lingüística y consultan bases de datos con información de este tipo. En la Figura 1, se detalla el funcionamiento de la síntesis del habla.

Figura 1
Sistema de conversión de texto en voz



Nota. Adaptado de *Estudio de la conversión de texto a voz basada em DNN: modelo base y fine tuning* (p. 3), por Peñas (2023).

Existen tres tipos principales de síntesis de voz: la síntesis por formantes, la síntesis articulatoria y la síntesis por concatenación. La síntesis por formantes produce el habla mediante la especificación previa de parámetros acústicos; la síntesis articulatoria lo hace a partir de parámetros que describen la posición y el movimiento de los articuladores, y la síntesis por concatenación crea el habla uniendo pequeños fragmentos de sonido para formar oraciones (Fernández, 2007).

2.3. Momentos espectrales de sonidos fricativos

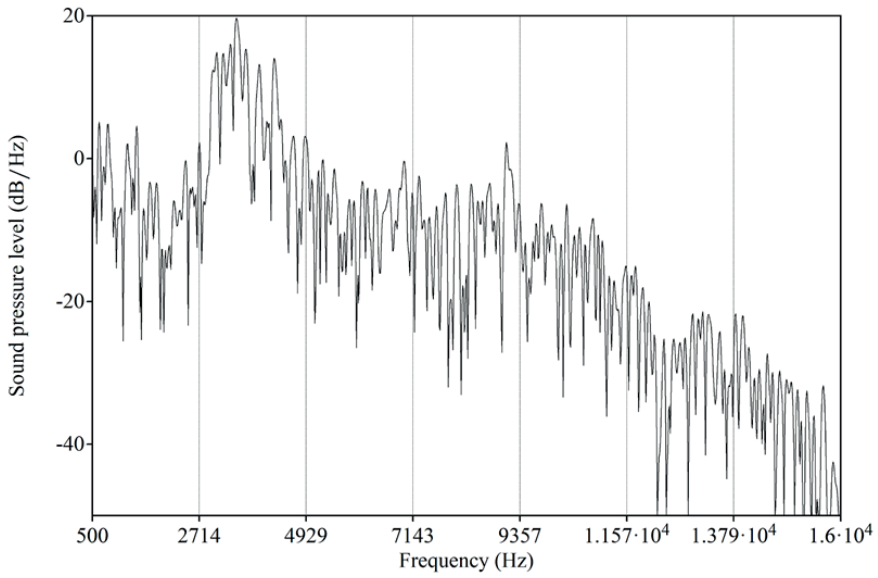
Las variables de análisis usadas en el estudio refieren a la forma del espectro. Específicamente, se analizan los cuatro momentos espectrales (centro de gravedad, desviación estándar, curtosis y asimetría) y el pico de máxima intensidad en el espectro LPC de la fricativa alveolar /s/.

En primer lugar, «el centro de gravedad de un espectro mide su concentración media de energía» (Cicres, 2011, p. 37); es decir, se obtiene al promediar todas las intensidades a lo largo de las frecuencias. Así, si hay una frecuencia en la que predomine mayor intensidad, lo más probable es que el centro de gravedad coincida o esté cercana a esta zona frecuencial. En el espectrograma, la frecuencia de este momento espectral se acerca al punto de mayor negror, asimismo, valores más altos indican una constricción más adelantada (Muñoz y Elvira-García, 2021).

Por su parte, «la desviación estándar (el segundo momento espectral) mide la distancia de las frecuencias del espectro respecto del centro de gravedad» (Cicres, 2011, p. 37). Así, indica si la energía se distribuye de forma más o menos aleatoria a lo largo de las frecuencias, es decir, si se observa todo el segmento de color gris o hay una parte blanca y otra muy intensa (Muñoz y Elvira-García, 2021). En inglés, la /θ/ tiene una desviación estándar mayor que /f/, mientras que /s/ posee una menor desviación estándar que /f/ (Tomiak, 1990, como se cita en Jongman *et al.*, 2000).

Como se observa en la Figura 1, el centro de gravedad coincide con la zona frecuencial donde se ubica el pico de mayor intensidad, lo que implica que su desviación estándar es baja, «puesto que el único pico de energía se corresponde con la zona del centro de gravedad» (Cicres, 2011, p. 37). Por otra parte, en la Figura 2, el centro de gravedad se encuentra en un punto intermedio entre las dos zonas frecuenciales que concentran los dos picos de mayor intensidad y, por eso mismo, tiene una desviación estándar mayor.

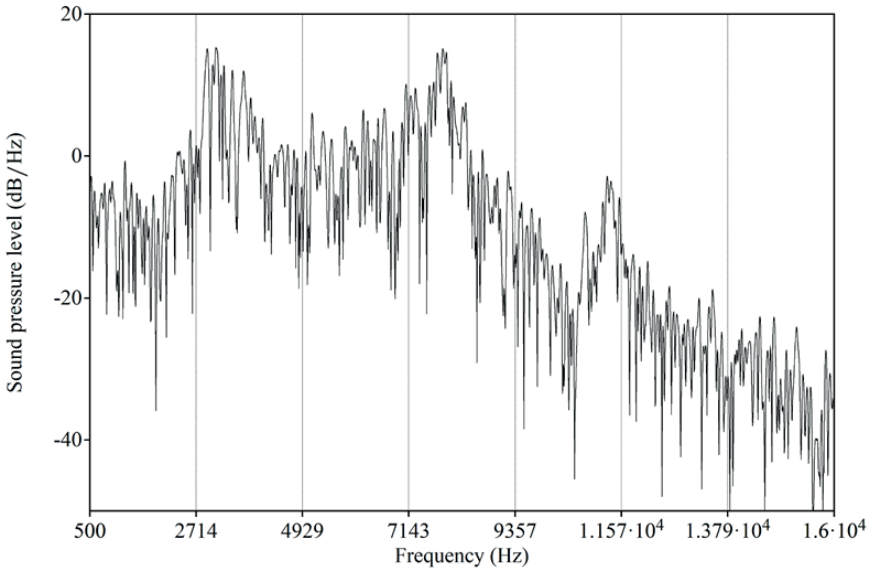
Figura 2
Espectro de la fricativa /s/ con centro de gravedad de 3572 Hz y desviación estándar de 1003 Hz



Nota. Elaboración propia

Figura 3

Espectro de la fricativa /s/ con centro de gravedad de 5210 Hz y desviación estándar de 2298 Hz

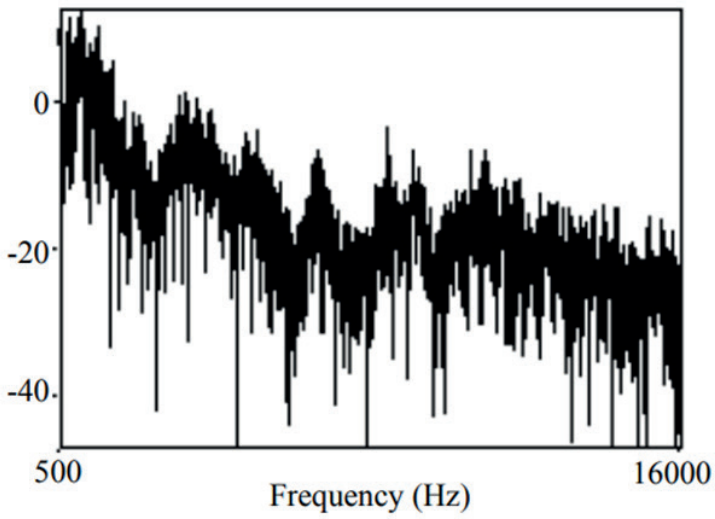


Nota. Elaboración propia

La curtosis (el tercer momento espectral) indica si la forma del espectro es puntiaguda, valores positivos reflejan espectros puntiagudos —Figura 4— y valores negativos indican distribuciones más planas del espectro —Figura 5— (Cicres, 2011). Así, «se esperan valores de coeficiente de curtosis más altos para las fricativas sibilantes y sobre todo para las [s] apicales y las variantes "silbadas" o de ultrafrecuencia» (Sadowsky y Perdomo, 2016, como se cita en Muñoz y Elvira-García, 2021, p. 803).

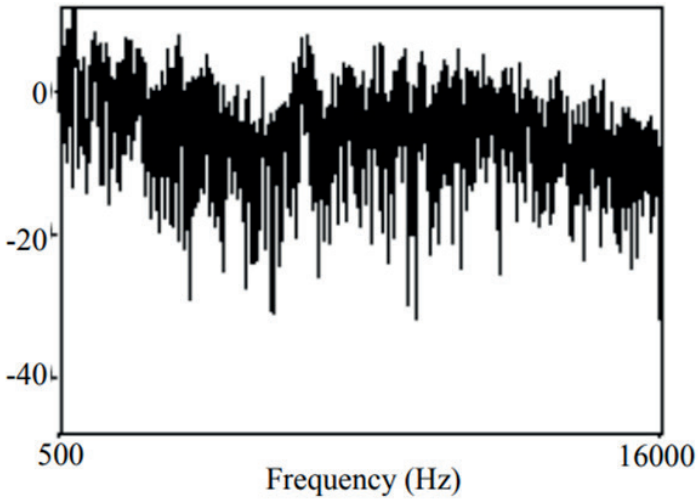
Figura 4

Espectro de una fricativa con curtosis positiva de 14,62



Nota. Tomado de *Los sonidos fricativos sordos y sus implicaciones forenses* (p. 39), por J. Cicres, 2011, Estudios Filológicos.

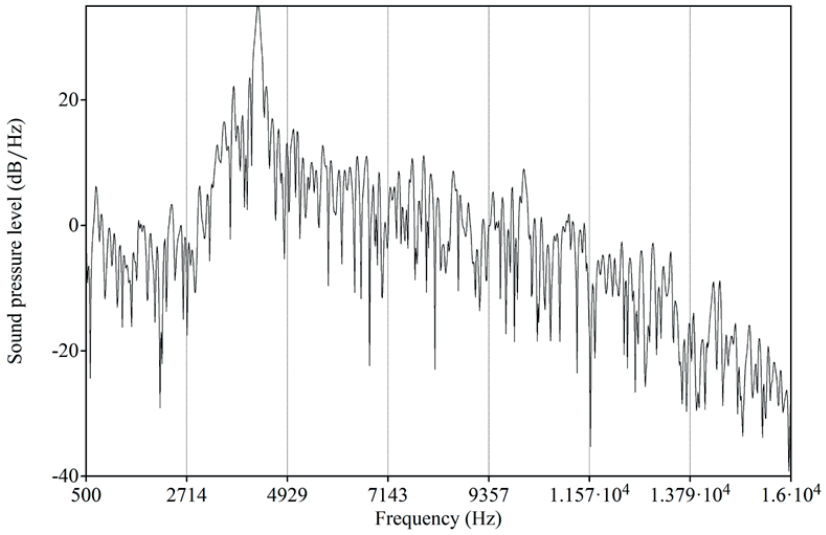
Figura 5
Espectro de una fricativa con curtosis negativa de –1,21



Nota. Tomado de *Los sonidos fricativos sordos y sus implicaciones forenses* (p. 39), por J. Cicres, 2011, Estudios Filológicos.

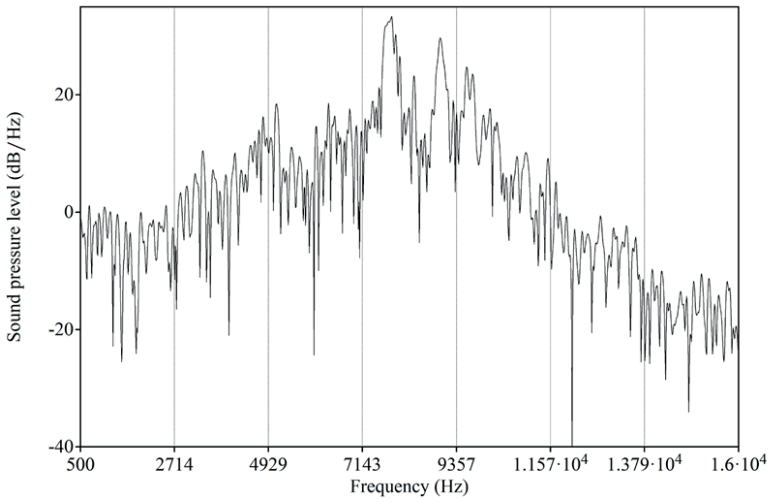
La asimetría mide la distribución de energía en ambos lados del centro de gravedad, valores positivos indican concentración de energía en las frecuencias bajas —Figura 6— y valores negativos indican espectros con concentración de energía en las frecuencias altas —Figura 7— (Cicres, 2011). Así, «se espera un coeficiente de asimetría más alto para las alveolares» (Muñoz y Elvira-García, 2021, p. 803).

Figura 6
Espectro de la fricativa /s/ con centro de gravedad de 4375 Hz y asimetría de 4,4



Nota. Elaboración propia

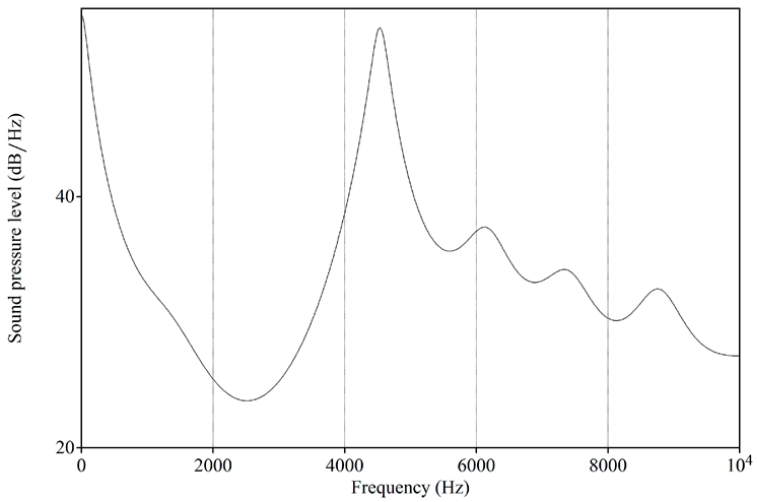
Figura 7
Espectro de la fricativa /s/ con centro de gravedad de 7871 Hz y asimetría de –1,09



Nota. Elaboración propia

Finalmente, en el espectro LPC aparecen destacados los picos de máxima intensidad, en fricativas, cada pico representa una zona frecuencial amplificada (Martínez y Fernández, 2013). La frecuencia del pico de máxima intensidad baja a medida que la constricción se mueve de adelante hacia atrás (Johnson, 2003), por lo que «articulatoriamente se ha planteado que fricativas más posteriores tienen el pico espectral a frecuencias más bajas» (Muñoz y Elvira-García, 2021, p. 802). Diferentes estudios aplican este parámetro para describir la variación alofónica de lenguas orales (Elías-Ulloa, 2011; Jimenez, 2021). Muñoz y Elvira-García (2021) reportan que el pico de máxima intensidad de la /s/ en coda del español de Antioquia (Colombia) se ubica en los 3791 Hz para varones. En la Figura 8, se observa el espectro LPC de una consonante fricativa alveolar sorda /s/.

Figura 8
Espectro LPC de la fricativa /s/ con el pico de máxima intensidad en los 4552 Hz



Nota. Elaboración propia

3. Metodología

Los estudios de alcance correlacional buscan determinar el grado de relación que existe entre dos o más variables en un contexto específico (Hernández *et al.*, 2014). En ese sentido, la presente investigación sigue ese alcance, porque se busca determinar si existe una relación entre las muestras de voz —artificial y natural— (variables independientes), y los cuatro momentos espectrales y el pico de máxima intensidad de /s/ en coda (variables dependientes).

Asimismo, el enfoque del estudio es cuantitativo porque, según Hernández *et al.* (2014), se enfoca en medir variables y analizar datos numéricos mediante métodos estadísticos con el objetivo de probar hipótesis —nula y alternativa— y evaluar las relaciones entre variables.

3.1. Recolección de datos

3.1.1. Recolección de datos de la voz natural

Para el presente estudio, se ha utilizado un corpus formado por un hablante masculino del español. El colaborador tiene veintitrés años y es natural del departamento de Lambayeque, provincia de Ferreñafe, reside en la ciudad de Lima metropolitana desde los dieciocho años y desde que llegó a Lima no ha frecuentado su ciudad natal. Asimismo, tiene educación secundaria completa y está cursando estudios técnicos. Para esta investigación, se han anonimizado los datos personales del colaborador.

El instrumento de recolección de datos consistió en la elicitación de un corpus de cien palabras (Anexo 1) en el que la fricativa alveolar /s/ se encontraba en posición de coda interna, en palabras bisilábicas —alguna de ellas inventadas—, en sílaba acentuada y ante las cinco vocales del español. De acuerdo con Martínez (1991), para estudiar algunos sonidos del habla en específico, se usa una frase cuyo nombre es *portadora* (alberga la palabra donde se encuentra el sonido objetivo) y esta debe seguir un esquema para todos los sonidos a recolectar. En ese sentido, las palabras del corpus fueron repetidas tres veces en la siguiente frase: «Yo digo la palabra _____ tres veces». Es necesario advertir que el uso de una frase portadora es para neutralizar factores prosódicos.

La grabación se realizó en un ambiente controlado, libre de ruidos externos que interfieran con la señal de grabación. Se utilizó una grabadora Zoom H5n pro y un micrófono tipo cardioide de marca Shure modelo WH20.

El resultado de las grabaciones consiste en cinco archivos de audio en formato WAV, los cuales se dividieron de acuerdo a la vocal que precede a la fricativa alveolar /s/, como se observa en la Tabla 1.

Tabla 1
Archivos de audio de la voz natural

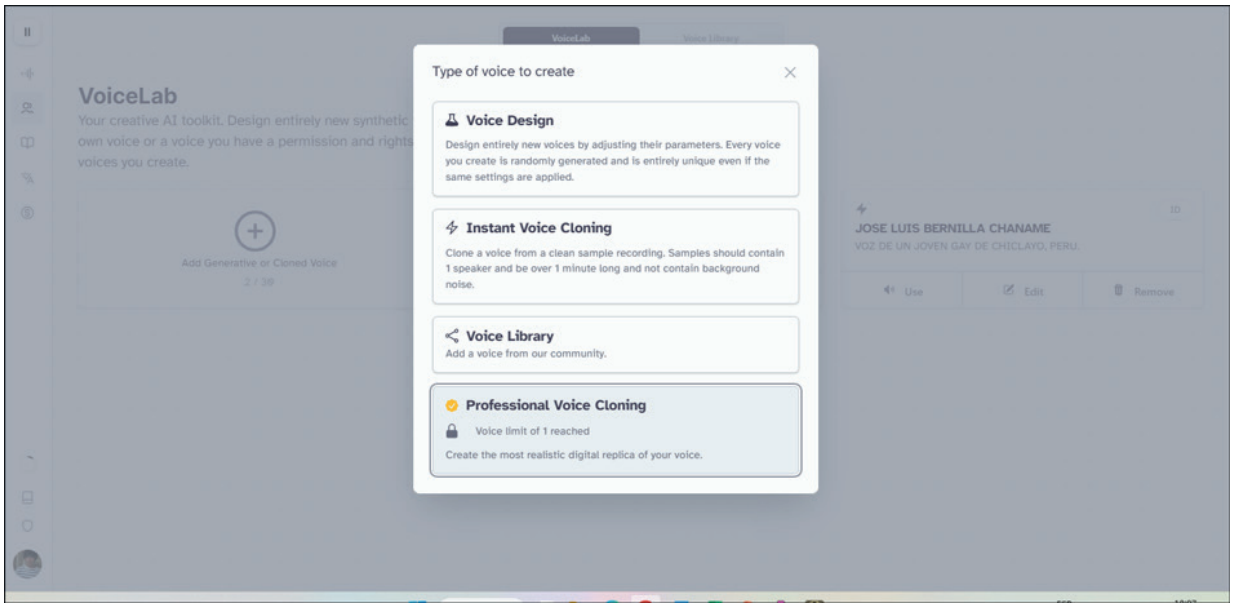
Nombre	Tamaño	Duración (min)
vocal_a.wav	14,858 KB	00:02:52
vocal_e.wav	13,984 KB	00:02:42
vocal_i.wav	14,033 KB	00:02:42
vocal_o.wav	14,367 KB	00:02:46
vocal_u.wav	13,075 KB	00:02:31

3.1.2. Recolección de datos de la voz artificial

Para generar la voz artificial, se utilizó el generador de voz con inteligencia artificial ElevenLabs, el cual permite clonar la voz humana con una cantidad determinada de *input*. ElevenLabs es una empresa de tecnología especializada en la creación de herramientas avanzadas de procesamiento del lenguaje natural (NLP) y síntesis de voz fundada en 2022. La compañía se enfoca en desarrollar un software de inteligencia artificial que pueda generar voz de alta calidad a partir de texto, a fin de crear voces realistas y naturales para diversas aplicaciones, como audiolibros, asistentes virtuales y otros contenidos multimedia. La tecnología de ElevenLabs utiliza modelos de aprendizaje profundo para ofrecer una síntesis de voz precisa y personalizable.

La interfaz de ElevenLabs se observa en la Figura 9, en ella se cargan los audios que van a servir para clonar la voz del colaborador. Una de las indicaciones de los desarrolladores consiste en tener material suficiente, aproximadamente más de treinta minutos solo de la voz objetivo para un mejor funcionamiento del proceso de clonado. Así, se utilizaron los audios provenientes de la recolección de datos de la voz natural para generar la voz artificial del colaborador.

Figura 9
Interfaz para clonar la voz del colaborador



Debido a que Elevenlabs necesita de un texto escrito, se usó el mismo corpus de cien palabras empleado para obtener la voz natural. De esta forma, se generaron cinco audios en formato *mp3*. Posteriormente, estos audios fueron convertidos a formato WAV; se mantuvo su frecuencia de muestreo, profundidad de bits y número de canales original —44100 Hz, 32 bits y formato mono, respectivamente— con el fin de ser procesados y analizados. Estos audios se presentan en la Tabla 2.

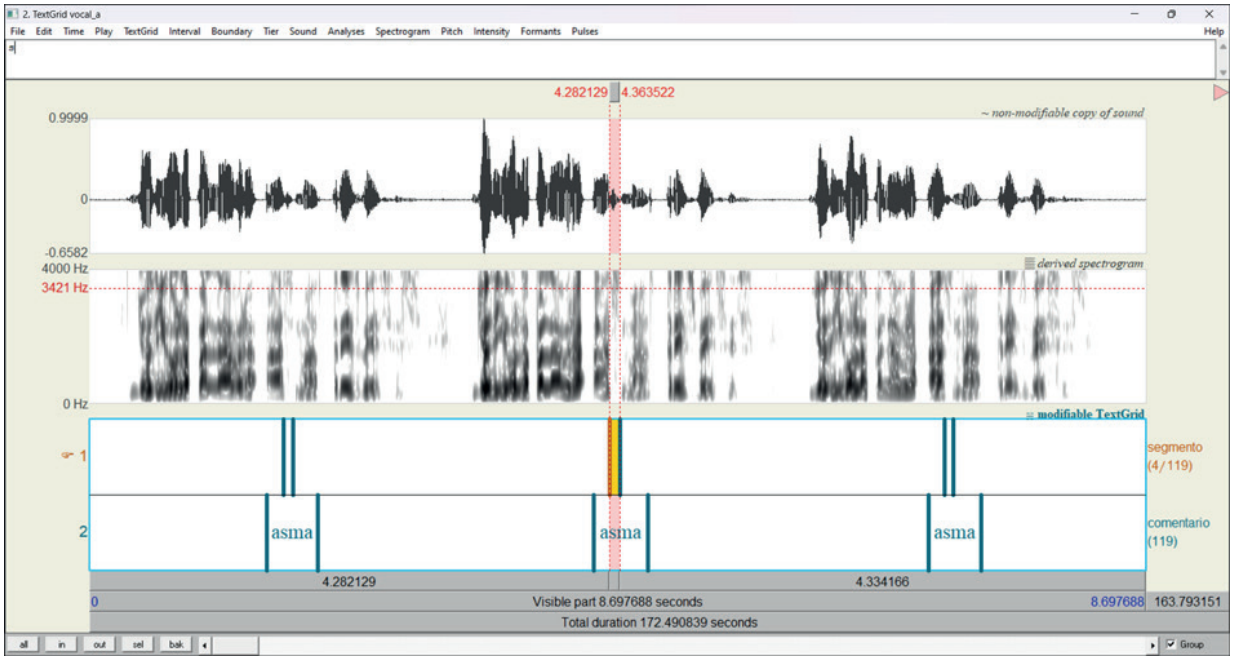
Tabla 2
Audios generados por la IA convertidos a wav

Nombre	Tamaño	Duración (min)
vocal_a.wav	16,024 KB	00:03:06
vocal_e.wav	15,693 KB	00:03:02
vocal_i.wav	15,443 KB	00:02:59
vocal_o.wav	15,569 KB	00:03:00
vocal_u.wav	15,585 KB	00:03:00

3.2. Extracción de datos acústicos

En Praat (Boersma y Weenink, 2024, versión 6.4.18), se crearon TextGrids para cada audio de ambas muestras (natural y artificial) con dos hileras. Luego, se procedió a segmentar los sonidos objetivo, lo cual ayuda a aislar el segmento sin interrupción de otros elementos conexos (Univaso, 2016). En la primera hilera, se segmenta a la fricativa alveolar /s/ y en la segunda hilera, a la palabra que contenía al segmento. A modo de ejemplo, en la Figura 10, se observa la segmentación de /s/ para la palabra *asma*.

Figura 10
Captura de pantalla de la segmentación en Praat



Los cuatro momentos espectrales se han extraído a partir del script *Zero-crossings-and-spectral-moments* v.2 de Elvira-García (2019). Los momentos espectrales se extrajeron a partir del espectro FFT de todo el segmento con un margen de 1 ms; asimismo, este *script* aplicó el filtro *pass Hann band* para acentuar las frecuencias de 1000-11000 y evitar los efectos que podría causar el F0 (Muñoz y Elvira-García, 2021).

Para extraer el pico de mayor intensidad, se usó el script Extractor de picos de Muñoz (2024), el cual aplicó el filtro *stop Hann band* para atenuar las frecuencias de 0-1000 y evitar ruidos ajenos a la fricativa. Finalmente, para dibujar los espectros LPC de la fricativa alveolar /s/, se usó el script Analisis_LPC_fricativas de Faucet (2024).

3.4. Análisis estadístico

En el propósito de cuantificar estadísticamente las diferencias entre los momentos espectrales y el pico de mayor intensidad de las variantes de cada muestra, se usó el análisis de varianza (ANOVA). El ANOVA se utilizó para analizar la influencia de la variable categórica (tipo de muestra: natural y artificial) en las variables cuantitativas (los momentos espectrales y el pico de mayor intensidad); esta prueba «determina si la variabilidad de los datos entre los diferentes grupos es mayor que en el interior de cada grupo» (Blecua, 2001, p. 85). Esta prueba considera una hipótesis nula y una hipótesis alternativa. La primera se da «si las medias poblacionales son iguales, [entonces] las medias muestrales de los diferentes grupos serán parecidas, existiendo entre ellas tan s[o]lo diferencias atribuibles al azar» (Pardo y Ruiz, 2005, p. 343). La segunda indica que hay al menos una diferencia significativa entre las medias muestrales de los diferentes grupos. Así, se acepta la hipótesis nula cuando el valor de la probabilidad (p) es mayor al nivel de significación —que en esta prueba fue 0.01— y se rechaza dicha hipótesis cuando el valor de p es menor al nivel de significación. Finalmente, es importante señalar que en este estudio se usó el ANOVA de un factor. Finalmente, los gráficos de caja de bigotes fueron creados a partir del programa Rstudio (versión 2023.12.1) (Rstudio Team, 2023).

4. Resultados

En este apartado, se examina la relación entre las variables del estudio, es decir, entre las muestras de voz artificial y natural, y los cuatro momentos espectrales y el pico de mayor intensidad de /s/. Para este propósito, se usó el análisis de varianza (ANOVA)[4] de un factor, el análisis indica que existen diferencias significativas (hipótesis alternativa) —$p<0.01$[5]— entre la voz natural y artificial en todos los parámetros. Esto significa que, aunque los parámetros analizados en las dos muestras de voz siempre mostrarán alguna diferencia entre ellos debido a que sus datos no son idénticos, una diferencia significativa indica que esta es suficientemente grande como para ser improbable que haya ocurrido por el azar.

4.1. Centro de Gravedad

En la Tabla 3, se presenta la frecuencia media del centro de gravedad de /s/ en la voz artificial y natural. Se puede observar que la frecuencia del centro de gravedad en la voz natural es más alta que en la voz artificial, lo que indica que la articulación de /s/ en la voz natural es ligeramente más adelantada.

4 El análisis de varianza (ANOVA) compara las medias de dos o más grupos y determina si al menos una de las medias es significativamente diferente de las demás.

5 El nivel de significación (α) de la prueba ANOVA fue de 0.01. Cuando el valor de p es inferior al nivel de significación existen diferencias significativas entre las muestras y cuando es superior no existen diferencias significativas.

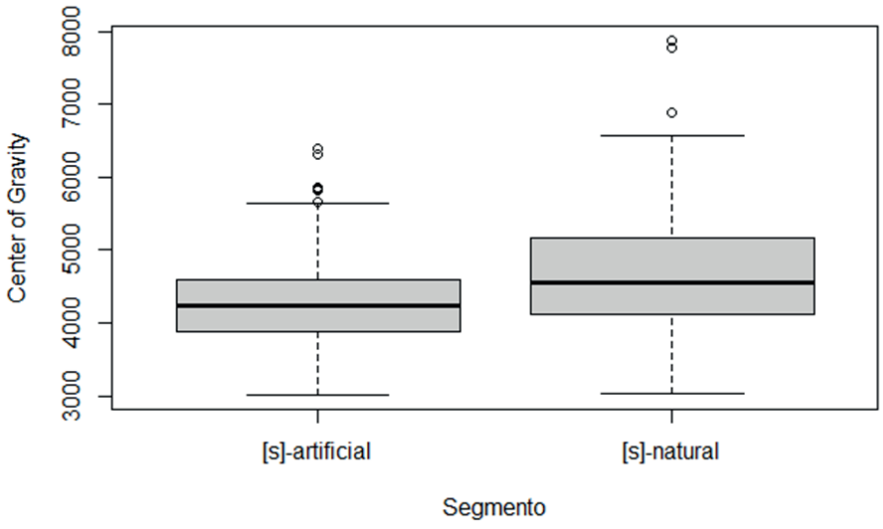
Tabla 3
Valores medios del centro de gravedad de /s/ en la voz artificial y natural

Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
a	59	(445.2) 4675	57	(813.1) 5338
e	58	(578.2) 4533	45	(578) 4578
i	56	(399.9) 4580	55	(612.9) 4952
o	59	(231.4) 3938	55	(312.4) 4248
u	60	(312) 3699	55	(665.9) 4191
Total	292	(566.4) 4279	267	(759.9) 4700

Nota. Frecuencia media (FM) y desviación estándar (DE).

En la Figura 11, se observa el diagrama de cajas de los valores correspondientes al centro de gravedad de /s/ en la voz artificial y natural.

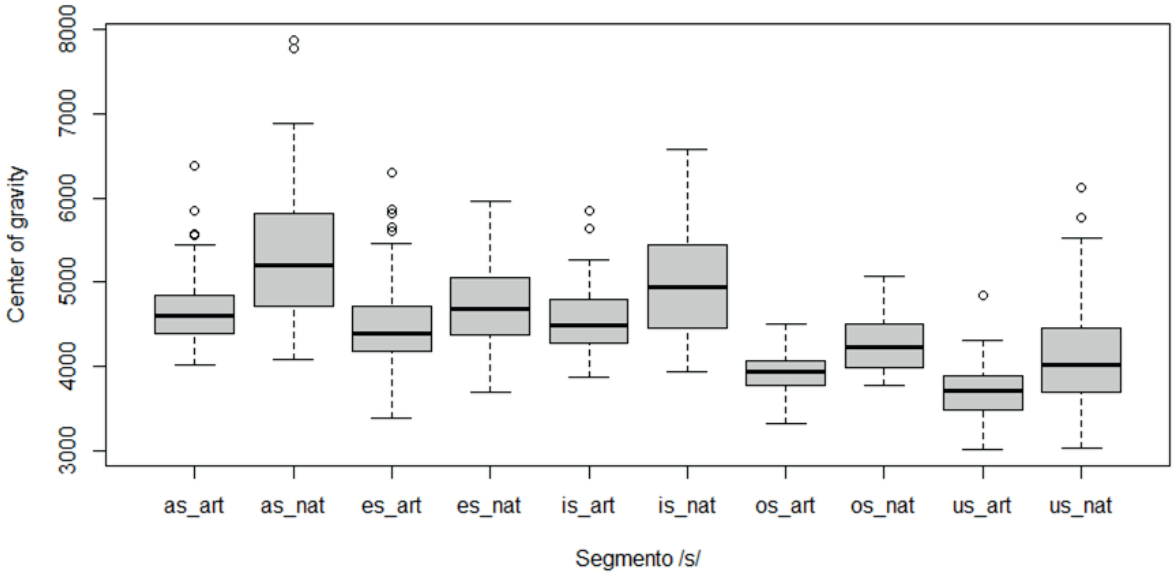
Figura 11
Diagrama de cajas del centro de gravedad de /s/ en la voz artificial y natural



El ANOVA de un factor indica que existen diferencias significativas entre las muestras de voz artificial y natural, y el centro de gravedad de /s/ ($p=3*10^{-13}$). Esto implica una relación entre ambas variables (el tipo de muestra de voz y el centro de gravedad).

Por su lado, en la Figura 12, se presenta el diagrama de cajas de la distribución de los valores del centro de gravedad de /s/ en la voz artificial y natural dividido por la vocal que le precede. Como se observa, el centro de gravedad de /s/ se ubica en frecuencias más bajas cuando le antecede una vocal posterior (/o/ y /u/) y los valores más altos se dan cuando le antecede la vocal /a/.

Figura 12
Diagrama de cajas del centro de gravedad de /s/ en la voz artificial y natural por vocales



Se realizaron pruebas de ANOVA de un factor para el centro de gravedad de /s/ agrupado por la vocal precedente tanto en la voz artificial y natural. Se observaron diferencias significativas entre las muestras de voz y el centro de gravedad de /s/ precedente a la mayoría de las vocales excepto con /e/ (a: $p=3*10^{-7}$, e: p=0.05, i: p=0.0002, o: $p=2*10^{-8}$ y u: $p=10^{-6}$).

4.2. Desviación estándar

En la Tabla 4, se presenta la frecuencia media de la desviación estándar de /s/ en la voz artificial y natural. Se observa una frecuencia media ligeramente más alta en la voz natural para este parámetro.

Tabla 4
Valores medios de la desviación estándar de /s/ en la voz artificial y natural

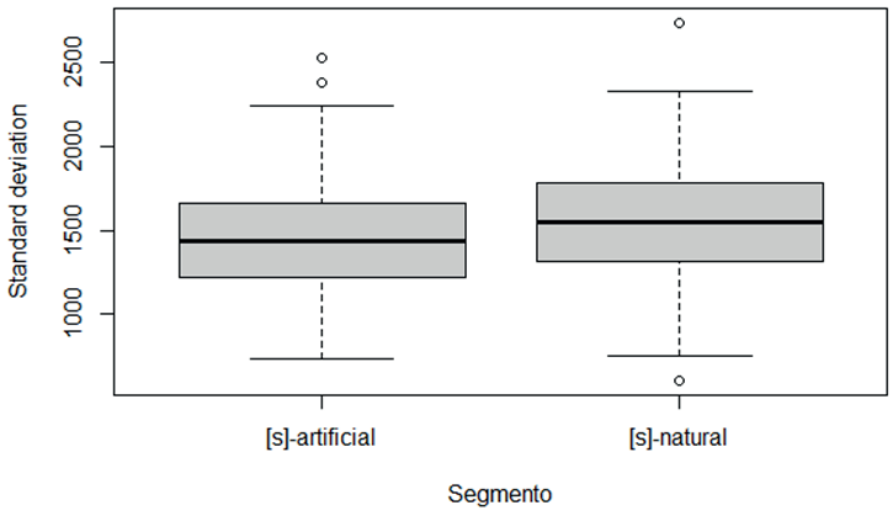
Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
a	59	(327.6) 1499	57	(318.2) 1498
e	58	(282.5) 1421	45	(340.6) 1508
i	56	(268.3) 1466	55	(390.6) 1444

Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
o	59	(282.3) 1256	55	(296.7) 1470
u	60	(311.6) 1590	55	(343.3) 1734
Total	292	(314) 1447	267	(352.8) 1532

Nota. Frecuencia media (FM) y desviación estándar (DE).

En la Figura 13, se observa la distribución de los valores correspondientes a la desviación estándar de /s/ en la voz artificial y natural. La desviación estándar de la voz natural es más alta para la voz natural, 1532 Hz respecto a los 1447 Hz de la voz artificial.

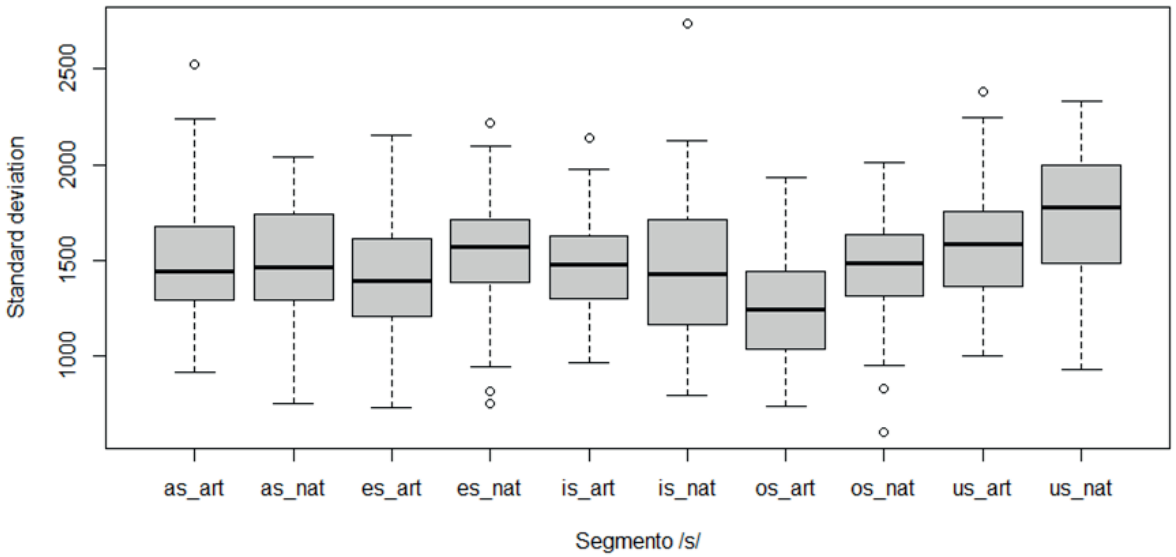
Figura 13
Diagrama de cajas de la desviación estándar de /s/ en la voz artificial y natural



El ANOVA de un factor indica que existen diferencias significativas entre las muestras de voz artificial y natural, y la desviación estándar de /s/ (p=0.002). Esto implica una relación entre ambas variables.

En la Figura 14, se presenta el diagrama de cajas de la distribución de los valores de la desviación estándar de /s/ en la voz artificial y natural dividido por la vocal precedente al segmento. Se observa que el valor más alto de la desviación estándar de /s/ se encuentra en la voz natural (1734 Hz) ante la vocal /u/. Esto se corrobora al observar su espectro LPC —ver Figura 21—, el cual presenta dos zonas frecuenciales que concentran los dos picos de mayor intensidad, lo que implica que la distancia de las frecuencias del espectro respecto del centro de gravedad sea mayor y la desviación estándar también.

Figura 14
Diagrama de cajas de la desviación estándar de /s/ en la voz artificial y natural por vocales



Se realizaron pruebas de ANOVA de un factor para la desviación estándar de /s/ agrupado por la vocal precedente tanto en la voz artificial y natural. Se observaron diferencias significativas entre las muestras de voz y la desviación estándar de /s/ únicamente cuando estuvo adyacente a la vocal /o/, en los demás contextos no se encontraron diferencias significativas (a: p=0.9, e: p=0.1, i: p=0.7, o: p=0.0001 y u: p=0.02).

4.3. Curtosis

En la Tabla 5, se presenta la frecuencia media de la curtosis de /s/ en la voz artificial y natural. Se puede observar que la frecuencia de dicho parámetro es más alta en la voz artificial, y se ubica alrededor de 5.8 y en 3.7 en la voz natural.

Tabla 5
Valores medios de la curtosis de /s/ en la voz artificial y natural

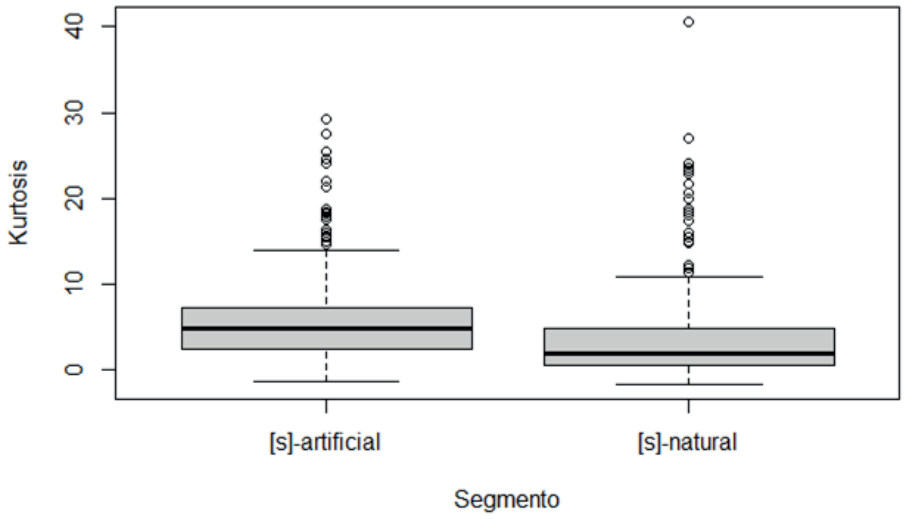
Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
a	59	(3) 3.6	57	(4.7) 3.3
e	58	(4.5) 4.3	45	(6.1) 4.3
i	56	(3.3) 4.1	55	(5.1) 4.4
o	59	(6.7) 10.6	55	(7.2) 5

Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
u	60	(4.4) 6.1	55	(4.6) 1.8
Total	292	(5.2) 5.8	267	(5.7) 3.7

Nota. Frecuencia media (FM) y desviación estándar (DE).

En la Figura 15, se observa la distribución de los valores correspondientes a la curtosis de /s/ en la voz artificial y natural. Se observa que en ambas muestras los valores de la curtosis fueron positivos, lo cual indica un predominio de espectros puntiagudos.

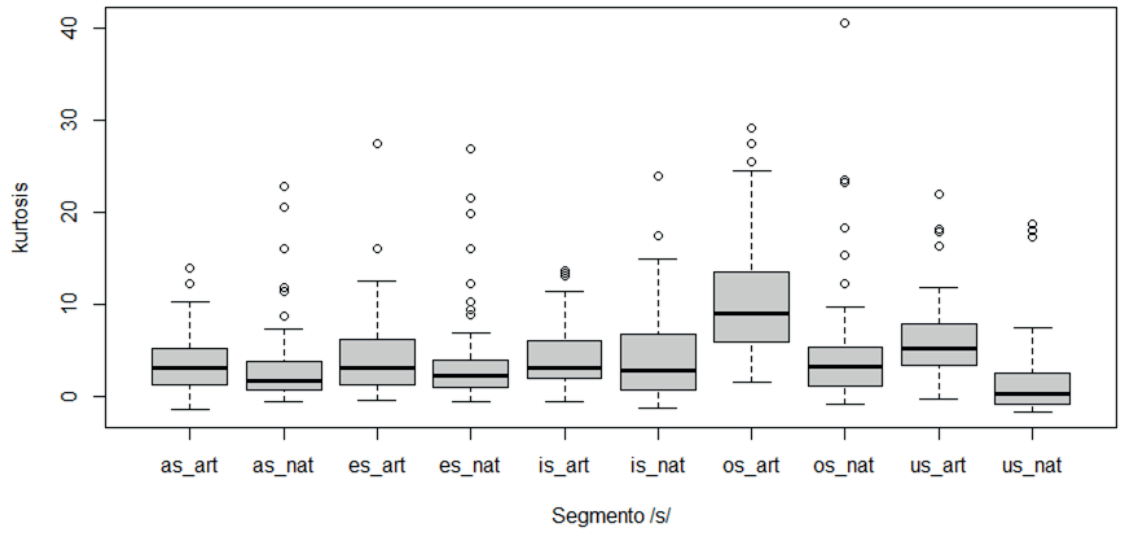
Figura 15
Diagrama de cajas de la curtosis de /s/ en la voz artificial y natural



El ANOVA de un factor indica que existen diferencias significativas entre las muestras de voz artificial y natural y la curtosis de /s/ (p=0.00002).

En la Figura 16, se presenta el diagrama de cajas de la distribución de los valores de la curtosis de /s/ en la voz artificial y natural dividido por la vocal que le precede. Se observa que el coeficiente de curtosis de /s/ más alto —el doble del promedio— se encuentra en la voz artificial ante la vocal /o/.

Figura 16
Diagrama de cajas de la curtosis de /s/ en la voz artificial y natural por vocales



Se realizaron pruebas de ANOVA de un factor para el coeficiente de curtosis de /s/ agrupado por la vocal precedente tanto en la voz artificial y natural. Se observaron diferencias significativas en las muestras de voz y la curtosis de /s/ únicamente ante las vocales /o/ y /u/; en los demás contextos, no se encontraron diferencias significativas (a: p=0.7, e: p=0.9, i: p=0.6, o: p=0.00004 y u: p=0.000002).

4.4. Asimetría

En la Tabla 6, se presenta la frecuencia media de la asimetría de /s/ en la voz artificial y natural. Se puede observar que la frecuencia de dicho parámetro es más alta en la voz artificial, y se ubica alrededor de 1.4 y en 1.2 en la voz natural.

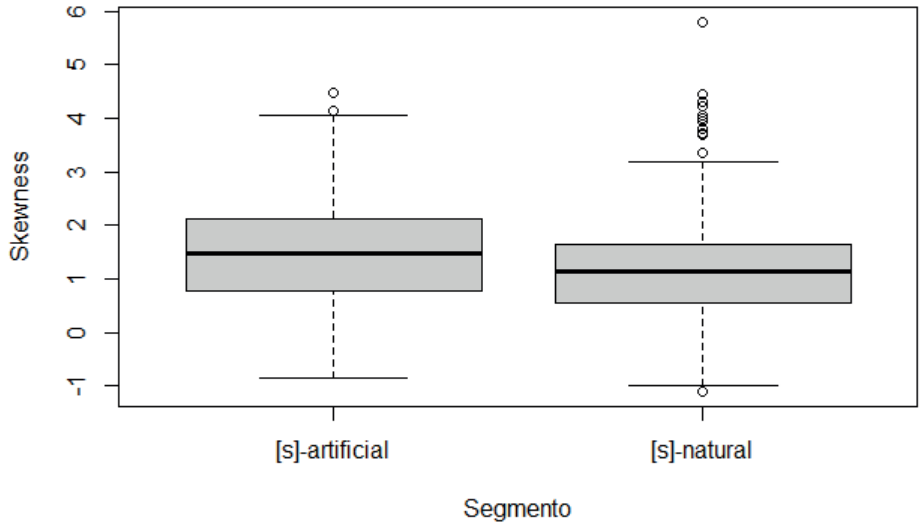
Tabla 6
Valores medios de la asimetría de /s/ en la voz artificial y natural

Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
a	59	(0.6) 0.8	57	(1) 0.6
e	58	(0.8) 0.8	45	(0.9) 1.1
i	56	(0.6) 1.2	55	(0.8) 1
o	59	(0.7) 2.4	55	(0.9) 1.8
u	60	(0.6) 2	55	(1) 1.2
Total	292	(0.9) 1.4	267	(1) 1.2

Nota. Frecuencia media (FM) y desviación estándar (DE).

En la Figura 17, se observa la distribución de los valores correspondientes a la asimetría de /s/ en la voz artificial y natural.

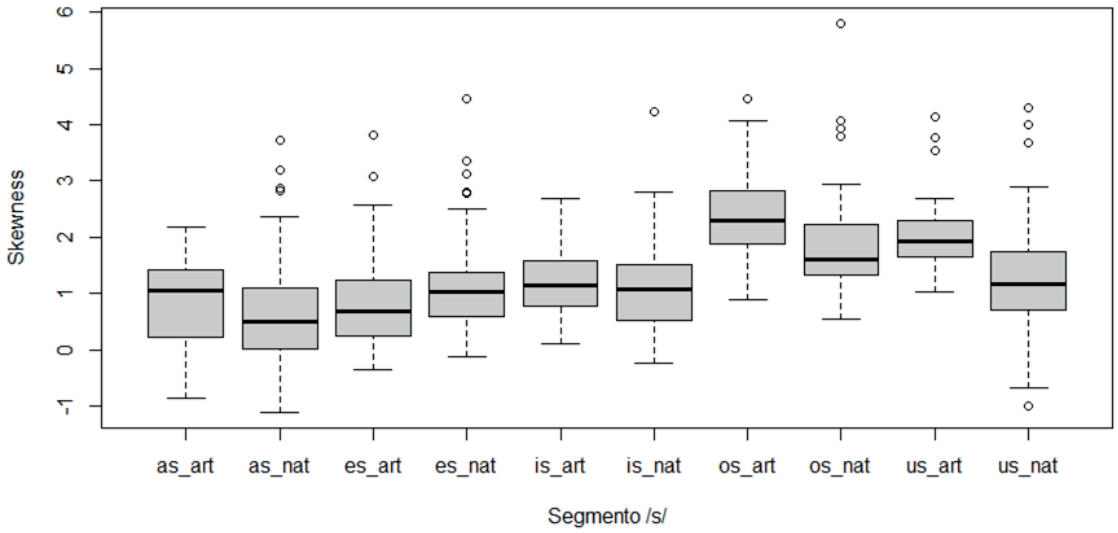
Figura 17
Diagrama de cajas de la asimetría de /s/ en la voz artificial y natural



El ANOVA de un factor indica que existen diferencias significativas entre las muestras de voz artificial y natural, y la asimetría de /s/ (p=0.001).

En la Figura 18, se presenta el diagrama de cajas de la distribución de los valores de la asimetría de /s/ en la voz artificial y natural dividido por la vocal que precede al segmento. Se observa que los valores más altos del coeficiente de asimetría se dan ante la vocal /o/ en ambas muestras.

Figura 18
Diagrama de cajas de la asimetría de /s/ en la voz artificial y natural por vocales



Por su parte, las pruebas de ANOVA de un factor para el coeficiente de asimetría de /s/ agrupado por la vocal adyacente tanto en la voz artificial y natural. Se observaron diferencias significativas entre las muestras de voz y la asimetría de /s/ únicamente cuando estuvo adyacente a la vocal /o/ y /u/, en los demás contextos no se encontraron diferencias significativas (a: p=0.2, e: p=0.03, i: p=0.3, o: p=0.0006 y u: p=0.00001).

4.5. Pico de máxima intensidad en espectros LPC

En la Tabla 7, se presenta la frecuencia media del pico de máxima intensidad de /s/ en la voz artificial y natural. Se puede observar que la frecuencia de dicho parámetro es más alta en la voz natural, por lo que la articulación de la fricativa en la voz natural es ligeramente más adelantada que en la voz artificial, lo cual se corresponde con lo descrito para el centro de gravedad.

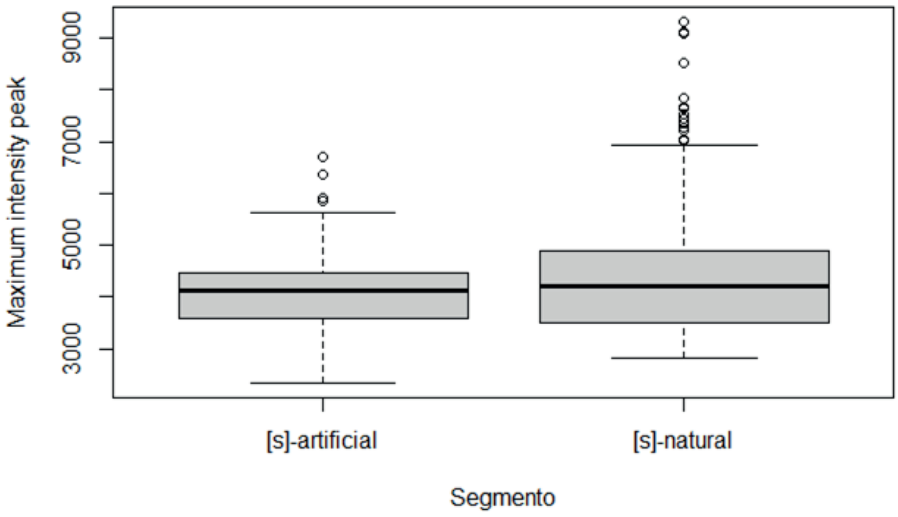
Tabla 7
Valores medios del pico de máxima intensidad de /s/ en la voz artificial y natural

Contexto	Voz artificial		Voz natural	
	Frecuencia de aparición	(DE) FM (Hz)	Frecuencia de aparición	(DE) FM (Hz)
a	53	(578.8) 4528	57	(1241.7) 5355
e	53	(374.1) 4472	45	(819.3) 4577
i	53	(492.2) 4390	55	(938.3) 4833
o	51	(246.4) 3635.8	55	(814.7) 3751
u	54	(298.6) 3320	55	(1262.6) 3522.7
Total	264	(647.2) 4070	267	(1248.4) 4408

Nota. Frecuencia media (FM) y desviación estándar (DE).

En la Figura 19, se observa la distribución de los valores correspondientes al pico de máxima intensidad de /s/ en la voz artificial y natural.

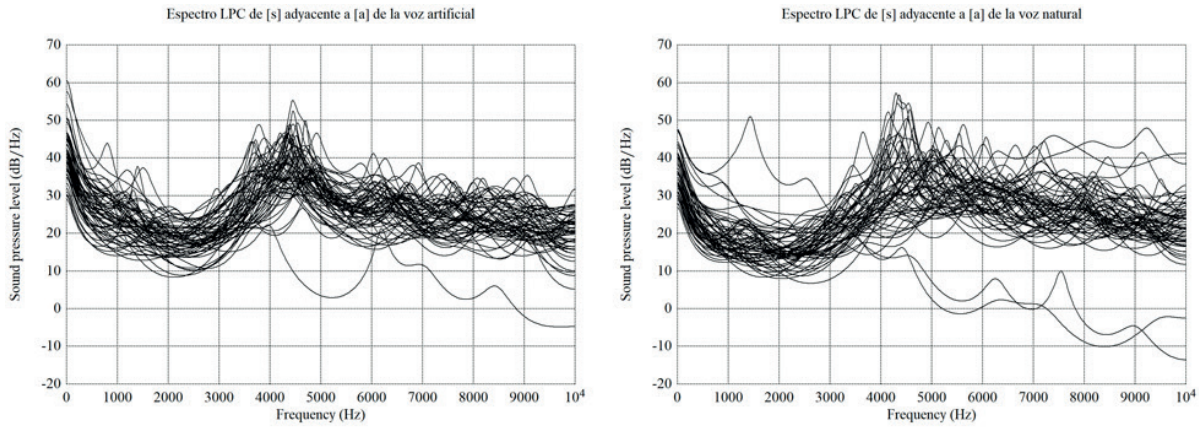
Figura 19
Diagrama de cajas del pico de máxima intensidad de /s/ en la voz artificial y natural

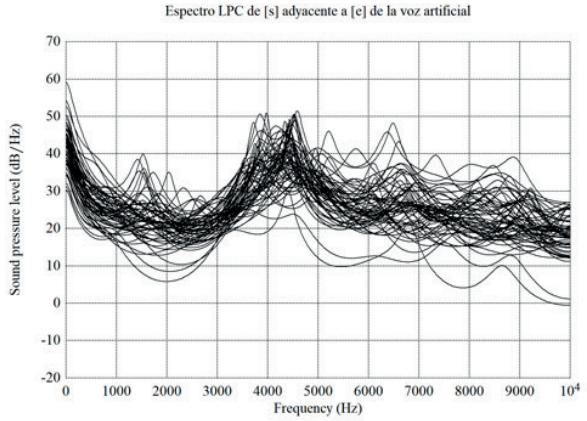


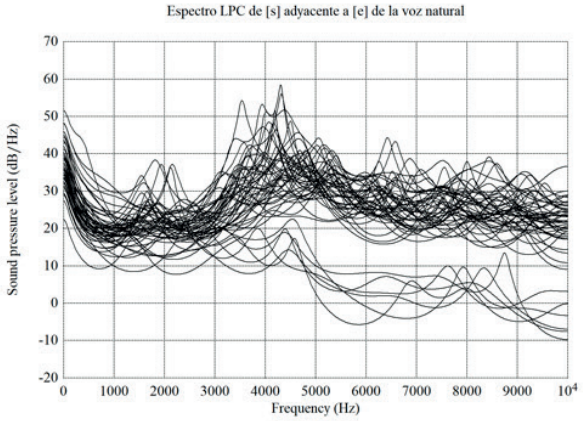
El ANOVA de un factor indica que existen diferencias significativas entre las muestras de voz artificial y natural, y el pico de máxima intensidad de /s/ (p=0.0001).

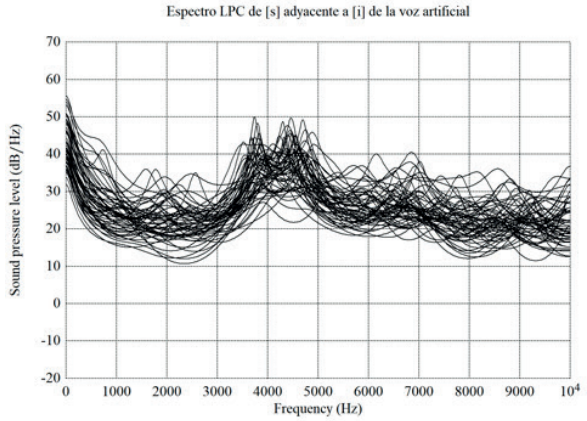
En la Figura 20, se presentan los espectros LPC de la fricativa alveolar /s/ de la voz natural y artificial dividido por la vocal que le precede. En ambas muestras, se observa que el pico de máxima intensidad con las frecuencias más altas se da ante la vocal /a/, mientras que los picos de mayor intensidad con frecuencias más bajas se observan ante las vocales posteriores /o/ y /u/, lo que sugiere que estas vocales podrían posteriorizar la articulación de /s/. Esto provocaría que el pico espectral de mayor intensidad se desplace hacia frecuencias más bajas

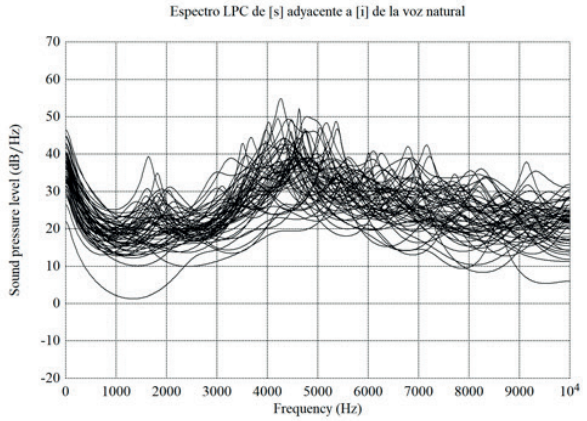
Figura 20
Espectros LPC de [s] de la voz artificial y natural por vocales

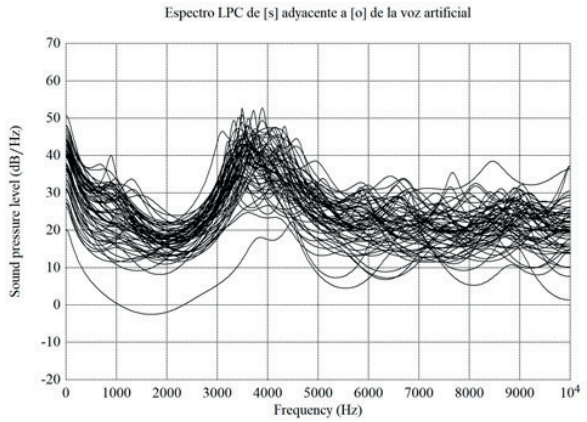


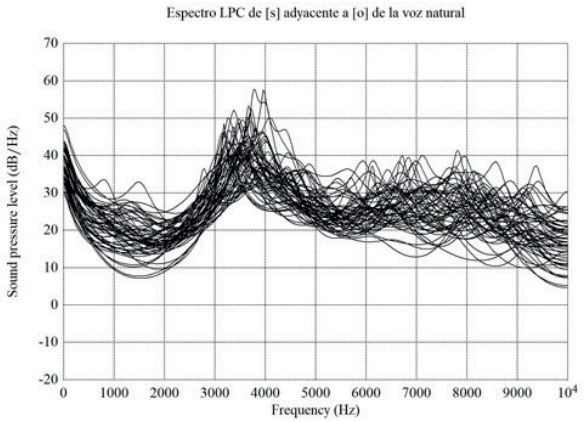

Espectro LPC de [s] adyacente a [e] de la voz artificial
Sound pressure level (dB/Hz)
Frequency (Hz)

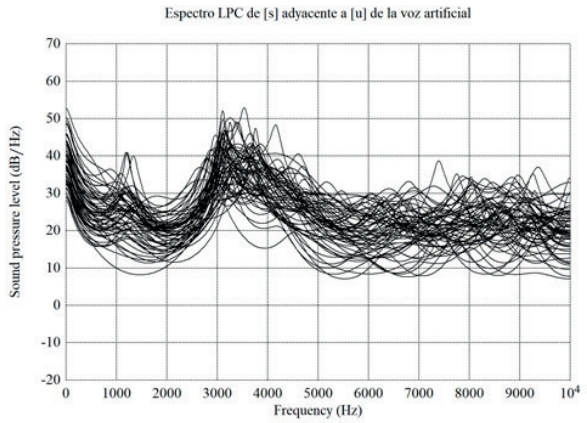

Espectro LPC de [s] adyacente a [e] de la voz natural
Sound pressure level (dB/Hz)
Frequency (Hz)

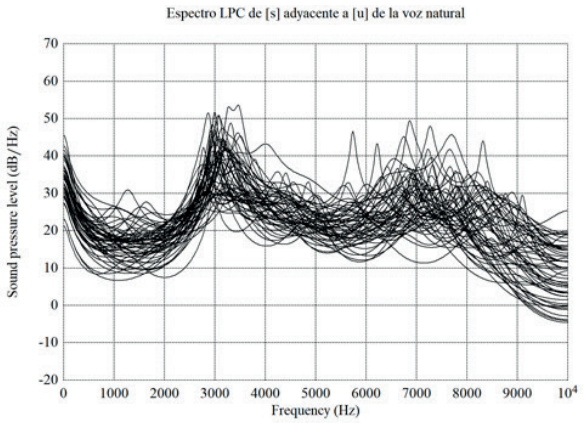

Espectro LPC de [s] adyacente a [i] de la voz artificial
Sound pressure level (dB/Hz)
Frequency (Hz)


Espectro LPC de [s] adyacente a [i] de la voz natural
Sound pressure level (dB/Hz)
Frequency (Hz)


Espectro LPC de [s] adyacente a [o] de la voz artificial
Sound pressure level (dB/Hz)
Frequency (Hz)

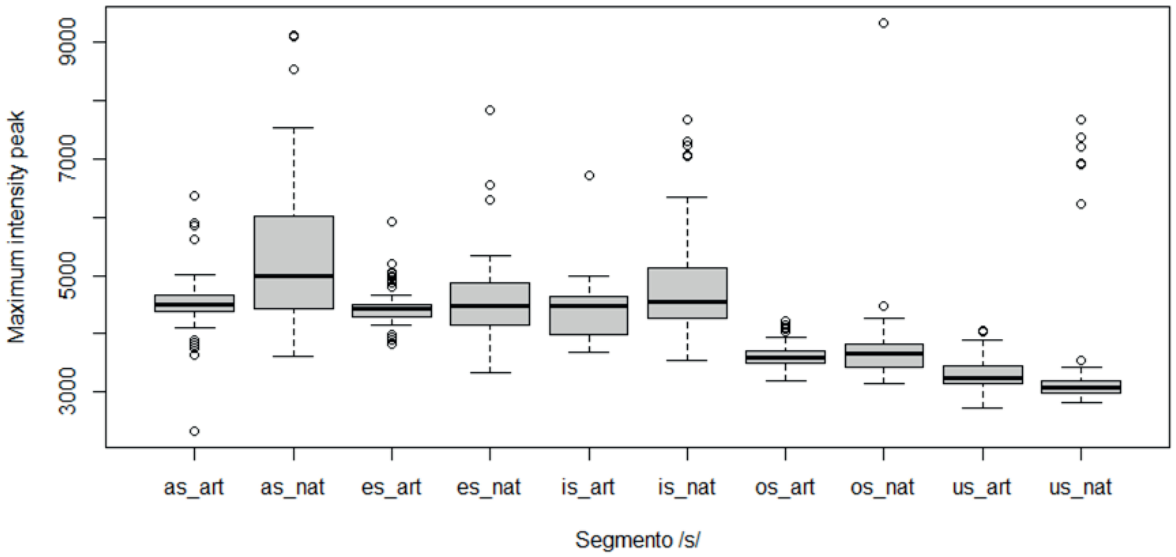

Espectro LPC de [s] adyacente a [o] de la voz natural
Sound pressure level (dB/Hz)
Frequency (Hz)


Espectro LPC de [s] adyacente a [u] de la voz artificial
Sound pressure level (dB/Hz)
Frequency (Hz)


Espectro LPC de [s] adyacente a [u] de la voz natural
Sound pressure level (dB/Hz)
Frequency (Hz)

En la Figura 21, se presenta el diagrama de cajas de la distribución de los valores del pico de máxima intensidad de /s/ en la voz artificial y natural agrupados por la vocal que precede al segmento.

Figura 21
Diagrama de cajas del pico de máxima intensidad de /s/ en la voz artificial y natural por vocales



Se realizaron pruebas de ANOVA de un factor para el pico de máxima intensidad de /s/ agrupados por la vocal adyacente tanto en la voz artificial y natural. Se observaron diferencias significativas entre las muestras de voz y el pico de máxima intensidad de /s/ únicamente cuando le precedió la vocal /a/ e /i/, en los demás contextos no se encontraron diferencias significativas (a: p=0.00002, e: p=0.4, i: p=0.002, o: p=0.3 y u: p=0.2).

5. Conclusiones

El objetivo de este estudio de caso ha sido comparar cinco parámetros acústicos (los cuatro momentos espectrales y el pico de mayor intensidad) de la fricativa alveolar sorda /s/ en posición de coda al interior de la palabra entre una muestra de voz natural y su contraparte clonada. Para ello, se analizaron más de quinientas realizaciones de [s] entre ambas muestras de voz.

Los resultados se dividieron en cinco apartados de acuerdo con el tipo de parámetros analizados como se observa en la Tabla 8. Los resultados indican que los cinco parámetros fueron útiles para diferenciar la voz natural de la voz artificial; no obstante, los resultados obtenidos al agrupar la fricativa alveolar /s/ según la vocal precedente presentan un panorama diferente. En ese sentido, se puede dividir los resultados en dos grupos: parámetros en donde el contexto vocálico no parece influir en la significancia de los resultados y parámetros en los que sí. En el primer grupo, se encuentra únicamente el centro de gravedad, considerado el parámetro más útil, que resultó significativo en cuatro contextos vocálicos: /a/, /i/, /o/ y /u/. En el segundo grupo, se puede hacer una subdivisión entre el pico de máxima intensidad, el cual fue significativo ante vocales no posteriores: /a/ e /i/ y la desviación estándar, la asimetría y la curtosis, los cuales fueron significativos ante vocales posteriores: /o/ y /u/.

Tabla 8
Resumen de los ANOVA de los parámetros acústicos evaluados

Parámetro	[as]		[es]		[is]		[os]		[us]		Global	
	p	**s.**	***p***	**s.**	***p***	**s.**	***p***	**s.**	***p***	**s.**	***p***	**s.**
Centro de gravedad	$3*10^{-7}$	Sí	0.05	No	$2*10^{-4}$	Sí	$2*10^{-8}$	Sí	10^{-6}	Sí	$3*10^{-13}$	Sí
Desviación estándar	0.9	No	0.1	No	0.7	No	10^{-4}	Sí	0.02	No	0.002	Sí
Asimetría	0.2	No	0.03	No	0.3	No	$6*10^{-4}$	Sí	10^{-5}	Sí	0.001	Sí
Curtosis	0.7	No	0.9	No	0.6	No	$4*10^{-5}$	Sí	$2*10^{-6}$	Sí	$2*10^{-5}$	Sí
Pico de máxima intensidad	$2*10^{-5}$	Sí	0.4	No	0.002	Sí	0.3	No	0.2	No	10^{-4}	Sí

Nota. Valor de p (p) y significancia (s).

El análisis de los parámetros acústicos de la fricativa /s/ según la vocal precedente sugiere patrones específicos relacionados con la articulación conjunta de estos sonidos. Mientras que algunos parámetros, como el centro de gravedad, parecen independientes del contexto vocálico, otros, como la desviación estándar, la asimetría y la curtosis, muestran una posible correlación articulatoria con las vocales posteriores (/o/ y /u/). No obstante, para confirmar estos hallazgos y comprender mejor estas dinámicas, sería necesario analizar un mayor número de muestras y considerar variaciones dialectales entre otros factores, esto permitiría evaluar si los patrones observados son consistentes en otros hablantes o si están influenciados por factores individuales o sociales.

En cuanto a las limitaciones del estudio, es importante señalar que los resultados provienen de datos obtenidos en laboratorio. Por ello, sería necesario realizar estudios similares en contextos de habla espontánea para contrastar estos hallazgos en situaciones probatorias; para ello, podrían emplearse llamadas telefónicas, grabaciones de audio o video como fuentes de análisis (Lazo y Rivas, 2022).

Finalmente, es fundamental analizar los parámetros utilizados en esta investigación dentro del enfoque de la razón de verosimilitud (*likelihood ratio*). Este enfoque requiere que las comparaciones se realicen utilizando poblaciones de referencia, lo que permitiría evaluar con mayor rigor la eficacia de cada parámetro acústico.

Referencias

Blecua, B. (2001). *Las vibrantes del español: manifestaciones acústicas y procesos fonéticos* [Tesis de doctorado, Universidad Autónoma de Barcelona]. Repositorio institucional de la Universidad Autónoma de Barcelona.

Boersma, P. y Weenink, D. (2024). *Praat: Doing Phonetics by Computer*. Versión 6.2.08. http://www.fon.hum.uva.nl/praat

Cicres, J. (2011). Los sonidos fricativos sordos y sus implicaciones forenses. *Estudios Filológicos*, *48*, 33-48.

Elías-Ulloa, E. (2011). *Una documentación acústica de la lengua shipibo-conibo (pano) (con un bosquejo fonológico)*. Fondo Editorial de la Pontificia Universidad Católica del Perú.

Faucet, C. (2024). *Herramientas para análisis acústico (plug-in para Praat)*. https://acortar.link/nQg8s5

Fernández, A. (2007). ¿Para qué sirve la Fonética? *Onomázein*, *15*(1), 39-51.

Hernández, R., Fernández, R. y Baptista, M. (2014). *Metodología de la investigación*. McGraw Hill.

Jimenez, J. (2021). Un estudio fonético-acústico preliminar sobre las consonantes nasales sordas de la lengua resígaro (arawak). *Lengua y Sociedad*, *20*(2), 295-312. https://revistasinvestigacion.unmsm.edu.pe/index.php/lenguaysociedad/article/view/22254/17827

Jimenez, J., Torres, F. y Cueva, O. (2022). Identificación de locutor a partir de la fonética forense: aplicación del software SplitsTree4 para una organización esquemática de los datos lingüísticos. *Boletín de la Academia Peruana de la Lengua*, *71*(71), 431-461. https://doi.org/10.46744/bapl.202201.014

Jimenez, J., Torres, F. y Cueva, O. (2024). Comparación forense de voces: un estudio preliminar sobre las diferencias entre una voz natural y una artificial para la investigación judicial. *Revista Oficial del Poder Judicial*, *16*(21), 53-81. https://revistas.pj.gob.pe/revista/index.php/ropj/article/view/881/1271

Johnson, K. (2003). *Acoustic and Auditory Phonetics*. (2.ª ed.). Blackwell Publishers.

Jongman, A., Wayland, R. y Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252-1263.

Lazo, V. y Rivas, G. (2022). La relación entre el extorsionador y la víctima en un caso de extorsión: una aproximación desde el análisis de la conversación. *Lengua y Sociedad*, *21*(2), 373-400. https://revistasinvestigacion.unmsm.edu.pe/index.php/lenguaysociedad/article/view/22535/18891

Llisterri, J., Carbó, C., Machuca, M., De la Mota, C., Riera, M. y Ríos, A. (2004). La conversión de texto en habla: aspectos lingüísticos. En M. Martí y J. Llisteri (eds.), *Tecnologías del texto y del habla* (pp. 145-186). Edicions de la Universitat de Barcelona – Fundación.

Martinez, E. (1991). *Fonética Experimental: Teoría y práctica*. Editorial Síntesis.

Martínez, E. y Fernández, A. (2013). *Manual de fonética española. Articulaciones y sonidos del español*. (2.ª ed.). Planeta.

Muñoz, D. y Elvira-García, W. (2021). Descripción acústica de la realización de [s] en Antioquia (Colombia). *Rilce. Revista de Filología Hispánica*, *37*(2), 793-818.

Muñoz, R. (2024). *Extractor de picos (plug-in para Praat)*. https://github.com/rolandomunoz

Pardo, A. y Ruiz, M. (2005). *Análisis de datos con SPSS 13 Base*. McGraw-Hill /Interamericana de España.

Peñas, I. (2023). *Estudio de la conversión de texto a voz basada en DNN: modelo base y fine-tuning* [Tesis de maestría, Universidad de Valladolid]. Repositorio institucional de la Universidad de Valladolid.

RStudio Team (2023). *RStudio: Integrated Development for R*. RStudio. http://www.rstudio.com/

San Segundo, E. (2023). *La fonética forense. Nuevos retos y nuevas líneas de investigación*. Octaedro.

San Segundo, E. y Delgado, J. (2024, 16 de octubre). *Deepfakes frente a voces de gemelos idénticos: un análisis acústico preliminar*. XIII Congreso Ibérico de Acústica, Faro-Portugal. https://doi.org/10.31219/osf.io/ckf6g

San Segundo, E. y Gibson, M. (2024, 16 de octubre). *Reconocimiento perceptivo de hablantes: un experimento con voces clonadas artificialmente y con voces de gemelos idénticos*. XIII Congreso Ibérico de Acústica, Faro-Portugal. https://doi.org/10.31219/osf.io/dtgwc

Univaso, P. (2016). *Identificación forense de hablantes en Argentina: un tutorial*. https://www.researchgate.net/publication/303616938_Identificacion_forense_de_hablantes_en_Argentina_un_tutorial

Anexos

Anexo 1. Palabras que conforman el instrumento de recolección de datos

	Coda dentro de palabra		**Total de palabras**
Núcleo vocálico			
a	asma aspa pasta caspa asno casto vasco gasco tasco chasco	Pasco rasca fiasco frasco lasca masca chasqui rasta casta raspa	20
e	pesco mezclo fresco pesto cesto crezco gresca testa cesta mesta	tesco flesco nesco plesco desco resco lesco tesco leslo pesmo	20
i	pista lista mismo risco triste isla ismo isma isco cisne	pisco misto tizna tisma quiste quispe chispa chisme chiste listo	20
o	poste posta mostro costo rostro ostra costra fosfo posdo nosdo	nosto oscar costa postre cosco rosca prosti crosti tosta tosti	20
u	lustre cusco busco busto mustia luspe muspe tuspe rusta rusco	busca cuspi tusco kusto puspa pusta pusca dusto duspi gusto	20
Cantidad de palabras			**100**

Contribución de los autores

Fernando Aaron Torres Castillo: análisis e interpretación de datos; concepción y diseño del trabajo; redacción y revisión crítica; aprobación final de la versión que se publicará. Se encargó de la verificación, ya sea como parte de la actividad o por separado, de la replicación/reproducibilidad general de los resultados/experimentos y otros resultados de investigación.

Oscar Esaul Cueva Sanchez: análisis e interpretación de datos; concepción y diseño del trabajo; redacción y revisión crítica; aprobación final de la versión que se publicará. Se encargó de la preparación, creación y/o presentación del trabajo publicado, específicamente, la visualización/ presentación de datos.

Jhon Jimenez Peña: análisis e interpretación de datos; concepción y diseño del trabajo; redacción y revisión crítica; aprobación final de la versión que se publicará. Se encargó de supervisar y liderar responsablemente la planificación y la ejecución de la actividad de investigación, incluyendo las tutorías externas.

Erika Amalec Shicshi Romero: análisis e interpretación de datos; concepción y diseño del trabajo; redacción y revisión crítica; aprobación final de la versión que se publicará. Se encargó de la preparación, creación y/o presentación del trabajo publicado, específicamente, la visualización/ presentación de datos.

Agradecimientos

Los autores expresan su agradecimiento, en primer lugar, a Akuma, quien permitió la clonación de su voz para esta investigación. En segundo lugar, a los revisores por sus valiosos aportes y observaciones. También a Maki por sus atinados comentarios sobre el tema de estudio y la naturaleza de la voz humana. Asimismo, a Tori por el significativo regalo que marcó su infancia, y al Dr. Tenma por ser un modelo de inspiración. Además, a los Lalos por su contribución del *script*. Finalmente, rinden homenaje a Satoru Gojo, a quien recuerdan con respeto y cariño.

Financiamiento

Autofinanciado

Conflicto de intereses

Los autores no presentan conflicto de interés.

Correspondencia: fernando.torresc@pucp.edu.pe

Trayectoria académica de los autores

Fernando Aaron Torres Castillo es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM) y maestro en Lingüística por la Pontificia Universidad Católica del Perú (PUCP). Sus intereses giran en torno al estudio de lenguas amerindias, entre ellas las familias quechua y arawak. Actualmente, labora como lingüista forense en la Oficina de Peritajes del Ministerio Público-Fiscalía de la Nación. Asimismo, es miembro adherente del grupo de investigación Kawsasun: Investigación intercultural para la formación docente y enseñanza de lenguas, del Instituto de Investigación de Lingüística Aplicada (CILA). También está adscrito como miembro del Gabinete de Lingüística Forense de la UNMSM. Pertenece al *Voice Deepfakes Project* financiado por el Ministerio de Ciencia e Innovación del Gobierno de España y la Unión Europea.

Oscar Esaúl Cueva Sanchez es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM). Sus intereses giran en torno a las áreas de fonética y análisis del discurso. Asimismo, es miembro del Gabinete de Lingüística Forense del Instituto de Investigación de Lingüística Aplicada (CILA). Pertenece al *Voice Deepfakes Project* financiado por el Ministerio de Ciencia e Innovación del Gobierno de España y la Unión Europea.

Jhon Jimenez Peña es licenciado en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM). Sus intereses están centrados en la fonética y la fonología de las lenguas originarias del Perú, con especial atención a la lengua arabela. Ha sido consultor en el Ministerio de Educación para la elaboración de fonologías que se han empleado en los procesos de normalización de alfabetos del arabela, el ocaina y el taushiro. También ha sido docente de los cursos de Fonología y Fonología Avanzada en el Curso Internacional de Lingüística, Traducción y Alfabetización (CILTA) del Instituto Lingüístico de Verano en los años 2018 a 2023, que se imparte en la Universidad Ricardo Palma. Además, ha sido expositor para el primer «Curso-Taller de fonética forense» organizado por el CILA-UNMSM. Es miembro del grupo de investigación Dolenper: Documentación lingüística de lenguas amenazadas en el Perú (CILA-UNMSM). Actualmente, labora como perito lingüista forense en la Oficina de Peritajes del Ministerio Público-Fiscalía de la Nación y es miembro del Gabinete de Lingüística Forense del CILA-UNMSM. Pertenece al *Voice Deepfakes Project* financiado por el Ministerio de Ciencia e Innovación del Gobierno de España y la Unión Europea.

Erika Amalec Shicshi Romero es bachiller en Lingüística por la Universidad Nacional Mayor de San Marcos (UNMSM). Sus intereses giran en torno al estudio de las lenguas originarias del Perú, con especial atención en el quechua. Actualmente, labora como lingüista forense en el Ministerio Público-Fiscalía de la Nación. Es miembro adherente del grupo de investigación Documentación Lingüística de Lenguas Amenazadas en el Perú (Dolenper) de la UNMSM.