DESCRIPCIÓN TRABAJO DE GRADO

AUTOR

Apellidos	Nombres
Morales Galindo	Andrés Mauricio

DIRECTOR (ES)

Apellidos	Nombres
Jiménez	Sergio

TRABAJO PARA OPTAR POR EL TÍTULO DE: Magister en Lingüística

TÍTULO DEL TRABAJO DE GRADO:
Patrones en los Factores Lingüísticos Alrededor de las Pausas Silenciosas en Entrevistas en el Español Panhispánico en YouTube

NOMBRE DEL PROGRAMA ACADÉMICO: Maestría en Lingüística

CIUDAD: Bogotá AÑO DE PRESENTACIÓN DEL TRABAJO: 2024

NÚMERO DE PÁGINAS: 52

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

MATERIAL ANEXO (Vídeo, audio, multimedia):

Duración del audiovisual: _____ Minutos.

Otro. ¿Cuál? _____

Sistema: Americano NTSC _____ Europeo PAL _____ SECAM _____

PREMIO O DISTINCIÓN (En caso de ser Laureadas o tener una mención especial):
__Trabajo laureado_Nota (4.9)_____

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**
Pausas silenciosas	Silent pause
Subtitulado sincronizado automático	Automatic synchronized subtitling
Detección de actividad de voz	Voice activity detection

RESUMEN DEL CONTENIDO Español (máximo 250 palabras):

Las pausas silenciosas en el habla espontánea permiten respirar, así como organizar, planear y evaluar el discurso. Sin embargo, su estudio se ha limitado a corpus relativamente pequeños, especialmente en el español, debido a restricciones asociadas con la recolección y anotación manual. Para superar esta limitación, se propone una nueva metodología de anotación automática de pausas para audios provenientes de videos de YouTube, aprovechando los avances en Inteligencia Artificial y procesamiento de señales en el campo de las comunicaciones. Esta metodología se aplicó en un corpus de 347 horas de entrevistas a celebridades del mundo hispanoparlante identificando aproximadamente 197.000 pausas. De este modo, se analizaron varios factores cognitivos y lingüísticos presentes en diferentes niveles de la lengua, encontrando distintos fenómenos alrededor de las pausas silenciosas y su duración. El estudio de dichos factores evidencia una relación entre la duración de las pausas y la complejidad del discurso a nivel cognitivo, tanto como con los rasgos en diferentes niveles de la lengua. Además, concluimos que la metodología de anotación automática propuesta es adecuada para la investigación lingüística.

RESUMEN DEL CONTENIDO Inglés (máximo 250 palabras):

The silent pauses in spontaneous speech allows to breath, as well to organize, plan and asses the discourse. However, it study has been limitated to relatively small corpus, specially in spanish due to restrictions asociated with manual annotation and recolection. In order to overcome this difficulty, it is proposed a new automatic annotation metodology of pauses for audio that came from YouTube videos, taking advantage of the advances in Artificial Intelligence and signal processing in the comunication field.This metodology was applied in a corpus of 347 hours of interviews to celebrities of the spanish-speaking world, identifying approximately 197.000 pauses. Thereby, it was analyzed several cognitive and linguistic patterns present in diferent language levels, finding several phenomena asociated to the silent pauses and its duration. The study of these factors shows a relationship between the pause duration and the complexity of the discourse at a cognitive level, as well as with the features at different levels of the language. Furthermore, we conclude that the proposed automatic annotation methodology is suitable for linguistic research.

BIBLIOTECA JOSÉ MANUEL RIVAS SACCONI

INFORMACION DEL TRABAJO DE GRADO

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

2. **Título del trabajo de grado: Patrones en los Factores Lingüísticos Alrededor de las Pausas Silenciosas en Entrevistas en el Español Panhispánico en YouTube**

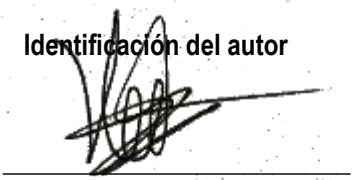
3. **Autoriza la consulta y publicación electrónica del trabajo de grado:**

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: Andrés Mauricio Morales Galindo

Documento de identidad: CC 1019127315 de Bogotá

PATRONES EN LOS FACTORES LINGÜÍSTICOS ALREDEDOR DE LAS PAUSAS SILENCIOSAS EN ENTREVISTAS EN EL ESPAÑOL PANHISPÁNICO EN YOUTUBE

Andrés Mauricio Morales Galindo

Artículo de investigación como modalidad de trabajo de grado

ASESOR
Sergio Jiménez

Instituto Caro y Cuervo
Facultad Seminario Andrés Bello
Maestría en Lingüística

TABLA DE CONTENIDO

Patrones en los Factores Lingüísticos Alrededor de las Pausas Silenciosas en Entrevistas en el Español Panhispánico en YouTube

Andrés Mauricio Morales Galindo

Resumen

Las pausas silenciosas en el habla espontánea permiten respirar, así como organizar, planear y evaluar el discurso. Sin embargo, su estudio se ha limitado a corpus relativamente pequeños, especialmente en el español, debido a restricciones asociadas con la recolección y anotación manual. Para superar esta limitación, se propone una nueva metodología de anotación automática de pausas para audios provenientes de videos de YouTube, aprovechando los avances en Inteligencia Artificial y procesamiento de señales en el campo de las comunicaciones. Esta metodología se aplicó en un corpus de 347 horas de entrevistas a celebridades del mundo hispanoparlante identificando aproximadamente 197.000 pausas. De este modo, se analizaron varios factores cognitivos y lingüísticos presentes en diferentes niveles de la lengua, encontrando distintos fenómenos alrededor de las pausas silenciosas y su duración. El estudio de dichos factores evidencia una relación entre la duración de las pausas y la complejidad del discurso a nivel cognitivo, tanto como con los rasgos en diferentes niveles de la lengua. Además, concluimos que la metodología de anotación automática propuesta es adecuada para la investigación lingüística.

Palabras clave: pausas silenciosas, subtitulado sincronizado automático, detección de actividad de voz

1. Introducción

Las pausas son un recurso lingüístico presente con frecuencia en los hablantes, en el cual se ve interrumpida la continuidad de una unidad comunicativa. Existen principalmente dos tipos de pausas, las silenciosas y las léxicas, también conocidas como pausas llenas. Stenström (1990) distingue entre las pausas silenciosas y las pausas léxicas argumentando que las primeras aparecen en partes estratégicas del enunciado y permiten a los hablantes organizar, planear y evaluar su discurso; mientras que, desde la perspectiva de Casalmiglia y Tusón (2001), las pausas léxicas se entienden como enunciados que no son planificados por los hablantes de manera

consciente, sino que revelan un alto grado de improvisación ya que aparecen a medida que el evento comunicativo va tomando lugar (puede tratarse de una conversación o de la lectura pública de un discurso, un texto, entre otros). Estudios más recientes, como el de Rose (2017), señalan que las pausas silenciosas presentan mayor asociación con los eventos de planeación y producción del habla en comparación con las pausas léxicas.

La presente investigación surge de la necesidad de hacer un estudio de las pausas silenciosas presentes en el ámbito hispanohablante en general, puesto que, al indagar la bibliografía relacionada se encuentra con estudios de índole experimental que elaboran el tratamiento de las pausas, tanto léxicas como no léxicas, de manera extensa y detallada en lengua inglesa, por ejemplo, el caso de Wang *et. al* (2010), Rochet Capellan y Fuchs (2014) o Crible *et. al* (2017), entre otros; pero con muy poco detalle en el español. En términos generales, se puede señalar que se hacen mediciones acústicas las cuales no ponen en consideración posibles factores fonológicos como las secuenciaciones de sílaba o de base computacional más común, como lo son las colocaciones (elementos adyacentes a la pausa), con los cuales se han podido establecer observaciones diatópicas y sociales en trabajos como el de Blondet (1999) o el de Pálvölgyi (2020).

Así mismo, las investigaciones encontradas hacen un tratamiento en diferentes lenguas, incluyendo la española con un corpus muy reducido y con alcances regionales restringidos (Borzi *et. al*, 2017) que, si bien puede constituirse como un importante punto de partida para evaluar las funciones y características de las pausas (Machuca, 2018), resultan poco idóneas para la generalización y tratamiento de los factores más recurrentes en una lengua ampliamente hablada como lo es el español en territorio hispanoamericano.

Ahora bien, ante la poca cantidad de trabajos publicados para el español, junto con los tamaños reducidos de corpus que dificultan observar los patrones presentes en el uso de pausas silenciosas en la lengua, surge la necesidad de explorar corpus de gran tamaño, los cuales generalmente no son utilizados debido a que los investigadores cuentan con tiempo y recursos limitados que dificultan una revisión más extensa de los datos. De hecho, en la investigación fonética experimental actual (Gries, 2009) se prefiere el uso de informantes y grabaciones recolectadas en ambientes controlados, ya que de este modo se puede tener buena calidad en los audios, además de la certeza del origen de los informantes y su relación con el fenómeno a estudiar.

Posteriormente, se opta por anotaciones de manera manual hechas por profesionales a través de herramientas computacionales como Praat (Boersma & Weenink, 2014), las cuales tienen costos y tiempos considerables.

Una posible alternativa ante esta dificultad es el uso de la identificación automática de pausas, la cual haría posible analizar corpus mucho más grandes para poder observar fenómenos difíciles de observar en corpus pequeños. Sin embargo, hasta ahora la identificación automática de pausas se ha considerado como una herramienta imprecisa cuyos resultados no son lo bastante certeros para ser utilizados en investigación lingüística (Lundholm, 2015, p.57).

Esta situación puede haber cambiado gracias a la combinación de los recientes desarrollos del área de Inteligencia Artificial en el campo de la Transcripción Automática (Lourador, 2023) y los avances consolidados en el área del procesamiento de señales en comunicaciones en el campo de Detección de Actividad de Voz (Giannakopoulos, 2009). La incorporación de estos avances haría que potencialmente mejore la calidad de la identificación automática.

Los audios recolectados en ambientes controlados garantizan su calidad, aunque en cierto sentido dificultan la recolección de corpus de gran extensión debido a los tiempos largos de grabación y la disponibilidad limitada de informantes. Además, los datos obtenidos no son del todo naturales ya que, como señaló Labov (1972), la presencia del investigador condiciona los resultados. No obstante, una perspectiva que se explora en esta investigación es utilizar una fuente de información de menor calidad, pero independiente del observador, que permita recolectar mayor cantidad de datos con el fin de identificar posibles fenómenos, y a su vez, obtener resultados más espontáneos. En ese sentido, se utilizó la plataforma YouTube debido a que provee gran cantidad de contenido de diálogos que se pueden analizar, a pesar de que son grabados con equipos propios y en condiciones heterogéneas; dicho de otro modo, la información obtenida con esta plataforma no garantiza una calidad de audio uniforme, pero sí es de fácil acceso, de gran extensión y de alguna manera más natural.

En YouTube el contenido audiovisual es bastante grande, con canales de todo tipo desde educativos realizados por organizaciones, hasta contenido de entretenimiento hecho por cualquier persona. Sin embargo,

para estudiar de manera más precisa las pausas silenciosas es preferible estudiar el habla informal en diálogos espontáneos, motivo por el cual se escoge como fuente de información las entrevistas realizadas a celebridades hablantes de español como lengua materna.

Por lo tanto, resulta relevante para el presente estudio, realizar un análisis de las pausas silenciosas y los patrones presentes en la duración, las colocaciones, la estructura silábica, así como otros rasgos presentes en el uso de las mismas, utilizando una cantidad considerable de entrevistas en videos de YouTube y realizando identificación automática de pausas.

Consecuentemente, en este estudio se presenta una nueva metodología para la detección automática de pausas silenciosas, la cual combina procesos de transcripción automática, alineación automática de transcripciones y audio, junto con métodos detección automática actividad de voz en audios. Esta metodología permitirá el análisis de un corpus de gran tamaño recolectado de YouTube, con el cual se espera identificar un número de pausas silenciosas hasta de tres órdenes de magnitud por encima de los estudios actuales.

El objetivo general de este estudio es investigar los factores lingüísticos presentes en el uso de las pausas silenciosas en el habla espontánea del español hispanoamericano, a gran escala, en la modalidad de entrevistas en YouTube. Para lograrlo, se plantean como objetivos específicos presentar una nueva metodología para la detección automática de pausas silenciosas en audios obtenidos de videos de YouTube; así como indagar cuáles son las relaciones entre los factores lingüísticos presentes alrededor de las pausas silenciosas con el proceso de elaboración del discurso de los hablantes.

La pregunta de investigación en este estudio es la siguiente: ¿Cuáles son los factores fonéticos, morfológicos, sintácticos, semánticos y cognitivos que se asocian con la caracterización de las pausas silenciosas del habla espontánea en español?

El resto del artículo se organiza de la siguiente manera: El marco conceptual se presenta en la sección 2 relacionado con las pausas silenciosas, la transcripción automática de audios, el subtitulado sincronizado automático y la detección de actividad de voz en señales. Posteriormente, en la sección 3 se muestran los

antecedentes de investigación. Luego, en la sección 4 se describe la metodología utilizada en este estudio para la detección automática de pausas silenciosas, lo cual conduce a la sección 5 donde se muestran los resultados obtenidos. La sección 6 está dedicada a la discusión de los hallazgos encontrados y, en último lugar, se presentan las conclusiones en la sección 7.

2. Marco Conceptual

2.1 Pausas Silenciosas en el Habla

Conviene señalar que diferentes autores han problematizado sobre cómo establecer diferencias frente al complejo tema de las pausas, así como sobre su definición. Para Cucchiarni, *et.al* (2002) las pausas silenciosas pueden ser entendidas como aquellos tramos de silencio que se encuentran presentes dentro de una señal de voz, al ser observada como un espectro de sonido de una grabación, mientras que Clark (2006) afirma que son una suspensión temporal en el ejercicio de comunicación de los hablantes, es decir, como aquellos elementos donde hubo una demora o interrupción en la continuidad de un enunciado. No obstante, vale la pena preguntarse por qué razón los hablantes realizan estos silencios. En primer lugar, es importante señalar que las pausas se producen como consecuencia de la naturaleza misma del aparato fonador, el cual requiere de la entrada y salida de aire para garantizar la emisión de sonidos. Para los hablantes no es posible continuar con su discurso si no inspiran y espiran el aire constantemente, ya que en los pulmones solo se puede almacenar cierta cantidad; de igual manera, se producen pausas como consecuencia de algún reflejo muscular de la laringe o al pasar saliva, lo que indica que son producto de una actividad fisiológica (Zellner, 1994).

De hecho, en el trabajo de Lundholm (2015) se afirma que las pausas se producen como resultado de dos eventos: el primero cuando ya no se tiene más aire en los pulmones, por lo que se requiere una inspiración; y el segundo, cuando ya no se sabe qué más decir, por lo que aparece una pausa que otorgue un tiempo extra para planear lo que va a decir a continuación. Desde esta perspectiva, se llega a la pausa o porque físicamente no es posible continuar con el enunciado, o porque cognitivamente se han agotado las ideas y es necesario revisar cómo continuar con la conversación.

Sin embargo, las pausas también se pueden entender como fragmentos temporales de interferencia producidos por los hablantes para organizar su discurso sin interrumpir completamente el enunciado, así como interrupciones de la continuidad del desarrollo expresivo del hablante que pueden revelar tanto alteraciones en el estado de ánimo -de seguro a dubitativo-, o también pueden ser entendidas como una estrategia para desarrollar nuevos marcadores discursivos. Por ejemplo, Ferreira (1991) encontró que si una oración tiene un sujeto y objeto sintácticamente complejos, los hablantes tienden a hacer una pausa entre el sujeto y la frase verbal con una duración incrementada. Esta autora también encontró que la complejidad sintáctica en la preparación del discurso tiene un efecto en el tiempo de pausa de iniciación de la producción oral.

La presencia de pausas silenciosas también puede indicar el dominio que tiene un hablante sobre una segunda lengua. Investigaciones como la de Housen y Kuiken (2009) junto al trabajo de Mora y Valls-Ferrer (2012) demuestran que la complejidad del discurso en términos léxicos y cognitivos, así como la fluidez y precisión guardan relación con la presencia de pausas silenciosas, puesto que estudiantes que participan en programas de intercambio hacen pausas más cortas en comparación con aquellos estudiantes con una formación tradicional, lo cual indica diferencias en la fluidez, precisión y complejidad en el uso de la segunda lengua. En esta misma línea, se destaca el trabajo de Williams (2023) quien se propone revisar la fluidez y eficiencia en el ejercicio de habla de una segunda lengua a partir de la presencia de seis elementos, que son: las pausas silenciosas, las pausas llenas, las repeticiones, prolongaciones, autocorrecciones y falsos arranques. En relación con el uso de pausas, este autor afirma que en la enseñanza de la segunda lengua se suele sugerir cambiar el uso de pausas silenciosas por pausas llenas para mejorar la fluidez en la conversación, lo cual puede funcionar para situaciones comunicativas sencillas como estar en una cafetería, pero que terminan siendo insuficientes para escenarios comunicativos más complejos como lo es la escuela o la discusión en seminarios, donde es indispensable buscar las palabras adecuadas y planear la organización sintáctica apropiada. Estos procesos cognitivos en situaciones donde se requieren tareas cognitivas más complejas resultan en su mayoría relacionados con el uso de pausas silenciosas (Williams, 2023, p.66).

Así mismo, las pausas pueden ser utilizadas por los hablantes como indicios para mantener o, incluso ceder, los turnos de conversación. Una pausa (léxica o no) puede ser utilizada para comunicar al interlocutor que el enunciado no está completo y que hay todavía información por señalar o, muy por el contrario, la pausa puede ser usada para indicar que no se desea agregar más información, permitiendo que el interlocutor intervenga y haga uso de su turno de conversación. Esta idea cobra sentido cuando se piensa en los trabajos relacionados con el análisis crítico del discurso (Schegloff ,1992), donde se observa detalladamente los efectos que tiene una pausa en el análisis exhaustivo de una conversación, así como la manera en qué puede variar la percepción de los hablantes si dicha pausa es más larga de lo esperado, llevando a inferencias y efectos que juegan un papel determinante en la construcción de sentido. Basta con pensar una conversación entre un hablante que realiza pausas más cortas y otro que realiza pausas más largas; desde la perspectiva del primer hablante la presencia de una pausa más larga le dará a entender que debe hacer uso de su turno de conversación y llenar el vacío que ha dejado su interlocutor, mientras que el segundo pensará que le están interrumpiendo y que no tienen interés en lo que está diciendo.

Otro buen ejemplo es la investigación de Tannen (1995), en la cual se pregunta por la valoración que hacen los hablantes de las ideas de otros a partir del estilo lingüístico propio, específicamente en lo relacionado al estilo que cada generó ha aprendido socialmente. Desde la perspectiva de esta autora, las pausas que hacen las mujeres suelen ser valoradas por parte de los hombres como muestra de duda e inseguridad, aunque un análisis más profundo permite ver que responde a los hábitos o estilos lingüísticos que se han aceptado como comunidad, ya que cuando el género femenino no hace pausas sino que presenta sus ideas como lo hacen los hombres, son valoradas entre ellas como autoritarias, poco empáticas y mandonas; a diferencia de los hombres que crecieron esperando ese liderazgo e intervención inmediata sin que sea considerado como algo negativo.

Para Sacks *et. al* (1974) es posible distinguir tres tipos de pausas en su función indicativa de turnos de conversación según la intención del hablante de ceder o no su turno. La primera sería la pausa propiamente dicha que es el silencio dentro del turno de conversación de un hablante y dentro del cual este mantiene su turno; la segunda es el salto o espacio que se produce cuando hay silencio dentro del discurso del primer hablante y el segundo hace uso de su turno sin que sea esta la intención del primer hablante; y la tercera es

el lapso que se produce cuando ninguno de los dos hablantes continúa con la conversación, por lo que la conversación se ve interrumpida de manera más larga, cortando así el flujo de la misma.

Por su parte, en el trabajo reciente de Curhan *et.al* (2022) se consideran dos aspectos funcionales de las pausas silenciosas en el contexto de negociación. Estos son la función de reflexión interna como una estrategia para crear valor al propio discurso y la percepción social como un mecanismo para reclamar valor e intimidar al interlocutor. El estudio realizado entre estudiantes universitarios de negocios en el noreste de Estados Unidos tanto en modalidad virtual como presencial permitió concluir que el silencio prolongado en la negociación tiene una fuerte relación con la creación de valor y la promoción de un ambiente deliberativo; aunque no se pudo demostrar que el uso de estos silencios tuviera un efecto negativo en la valoración del discurso de la contraparte.

Desde los años sesenta, las pausas se han reconocido como marcadores para la caracterización e identificación de desórdenes en la capacidad cognitiva de comunicación de los hablantes (Schegloff, 2003). Dentro de los trabajos recientes se destaca el de Potagas *et.al* (2022) quienes aseguran que la presencia de pausas silenciosas puede ser utilizada como un biomarcador para detectar afasia progresiva en etapas iniciales; junto con el de Balogh, *et.al* (2023) en el cual se propone que el análisis de la fluidez lingüística, la presencia de dudas, pausas silenciosas, así como repeticiones incoherentes en el discurso de hablantes puede ayudar a detectar, a través de herramientas automáticas, los posibles deterioros cognitivos leves en las personas.

2.2 Transcripción Automática de Audios

Dentro de los avances en Inteligencia Artificial más recientes, particularmente los métodos basados en Aprendizaje Profundo (*Deep Learning*), se encuentra la automatización de tareas relacionadas con el lenguaje con precisión similar o superior a la de un ser humano, siendo la transcripción de audios una de ellas, conocida como *Automatic Speech Recognition* (ASR). Actualmente, se considera que una de las mejores herramientas para transcripción automática es Whisper (Radford *et al.*, 2023), la cual está disponible como una librería para el lenguaje Python y es de uso libre. Si bien el propósito principal de esta herramienta no es la investigación lingüística, ha sido ampliamente utilizada en investigación a partir de su lanzamiento. Esta

investigación no es la excepción, puesto que fue utilizada para transcribir el corpus de entrevistas a celebridades.

2.3 Subtitulado Sincronizado Automático

No obstante, una de las limitaciones de la transcripción automática de audios es que genera las transcripciones, pero no entrega marcas de tiempo que permitan conocer en qué momento del video (de manera exacta) fue pronunciada y detectada una palabra, lo cual es necesario para realizar el análisis de las pausas. En ese sentido, la tarea de ubicar las palabras en los videos a medida que son pronunciadas es conocida como *Synchronized Captioning,* que ha sido usado en investigación para mejorar procesos de aprendizaje de segunda lengua (Mirzaei, *et. al.* ,2017). Esta herramienta hace más atractivos los videos, además de hacerlos inclusivos para personas con discapacidades auditivas o visuales, junto con usuarios de lengua extranjera.

Otra tecnología de reciente aparición, junto con el auge de plataformas de videos cortos como Tiktok, es el Subtitulado Sincronizado Automático (*Automatic Sychronized Captioning, ASC*) (Martín, *et al.*, 2021) que consiste en determinar automáticamente en qué momento en el audio fueron pronunciadas las palabras, con el fin presentarlas en el video exactamente cuando están siendo pronunciadas de manera sincronizada con el audio, el video y la transcripción. Este proceso se realiza sin intervención humana alineando la transcripción textual obtenida automáticamente con la señal de audio utilizando un algoritmo de alineamiento. Una herramienta para este propósito, como Whisper-timestamped (Louradour, 2023), que al igual que Whisper es una librería libre para Python, usa al mismo Whisper para obtener las transcripciones, y *Dynamic-Time-Warping* (Giorgino, 2009) con el fin de alinear las palabras con marcas de tiempo en la señal de audio.

2.4 Detección de Actividad de Voz en Señales

Uno de los mayores desafíos en la presente investigación es la identificación de las pausas puesto que la naturaleza de los videos escogidos presenta la interacción entre dos hablantes, así como sonidos de fondo, música, o incluso de la presentación de la entrevista. Es decir, se necesita reconocer con precisión dónde ocurren diálogos y dónde, a pesar de existir sonido, no hubo conversación; por tanto tampoco hubo pausas.

Por esta razón, se recurre al método propuesto por Giannakopoulos (2009), el cual permite diferenciar claramente los segmentos en una señal de audio que contienen sonido de aquellos segmentos donde hay silencio. Si bien el objetivo de este investigador es detectar los segmentos donde hay sonido para poder analizar los enunciados sin la presencia de ruido de fondo o de silencio, es posible utilizar este método para identificar, con mayor seguridad, dónde se encuentran los silencios en las señales de audio. De este modo, es posible analizar cuáles son las colocaciones más frecuentes en presencia de pausas silenciosas ya que en lugar de extraer los silencios, lo que se hace es evaluar los elementos lingüísticos adyacentes más frecuentes junto con sus respectivas características.

Para Giannakopoulos, los silencios en las señales de audio se pueden identificar a través de la energía de la señal y el centro de gravedad espectral calculados en segmentos de la misma señal. Para comprender mejor este enfoque, basta con pensar que en una señal de audio ocurre un silencio sin habla en aquellos tramos donde, tanto la energía de la señal (intensidad), como las frecuencias de los sonidos en el tramo son bajos, lo cual se confirma al observar que generalmente el ruido de fondo en el audio presenta frecuencias bajas.

Para lograr identificar esas bajas frecuencias, en primer lugar el audio se divide en segmentos cortos de igual duración que permitan comparar la información de manera sencilla. La energía es entendida como la intensidad de la señal en el segmento de audio estudiado. Esto significa que la señal de audio representada como una lista de números (muestras) se divide en segmentos de tamaño N muestras. Por ejemplo, un segmento en una señal de audio que fue obtenida a una rata de muestreo de 22.050Hz, o sea 22.050 muestras por cada segundo, podría tener un largo del segmento adecuado de 50 milisegundos, lo que correspondería a que cada segmento tenga 1.102 muestras. A cada uno de estos segmentos de tamaño fijo se le calcula la energía de la señal elevando al cuadrado los valores numéricos de las muestras y promediando estos valores. De esta manera se obtiene un valor de la energía de la señal para cada uno de los segmentos, lo cual produce la función de energía de la señal. Así, aquellos segmentos con una intensidad de la señal baja, es decir, de bajo volumen sonoro, tienen un valor de la energía bajo, en contraste con segmentos donde ocurre habla humana audible, puesto que la energía de la señal en estos casos obtiene un valor alto. Esto de manera formal corresponde a la siguiente ecuación:

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} \quad x_i(n)^2$$

Donde N es el número de muestras de los segmentos en los que se divide la señal y cada una las muestras del i-ésimo segmento de la señal se representa por $x_i(n)$.

Por otra parte, el centro de gravedad espectral permite obtener un valor numérico que representa la frecuencia predominante de los sonidos en cada segmento. Para esto, se aplica la transformada de Fourier discreta que convierte el segmento de audio del dominio del tiempo al dominio de la frecuencia. Posteriormente, a esta transformación espectral del segmento se le obtiene la frecuencia predominante usando una analogía física, que es la determinación del centro de gravedad (o de masa) de la representación gráfica de la transformación.

En la Figura 1, se ilustran dos segmentos de audio simulado con su transformación espectral y la ubicación del centro de gravedad en cada uno. En el segmento de la parte superior se observa que en él predominan las frecuencias bajas y en el inferior predominan las frecuencias altas. Por este motivo, el centro de gravedad representado por el punto de apoyo triangular de color rojo, se encuentra en una frecuencia baja para el segmento superior, a diferencia de la frecuencia alta en el inferior. El centro de gravedad sería el punto donde la gráfica estaría en equilibrio si las barras de los componentes espectrales estuviesen hechas de un material uniforme con peso y masa. En este ejemplo la gráfica superior podría considerarse como ruido de fondo por su bajo centro de gravedad, y la inferior como posible habla humana.

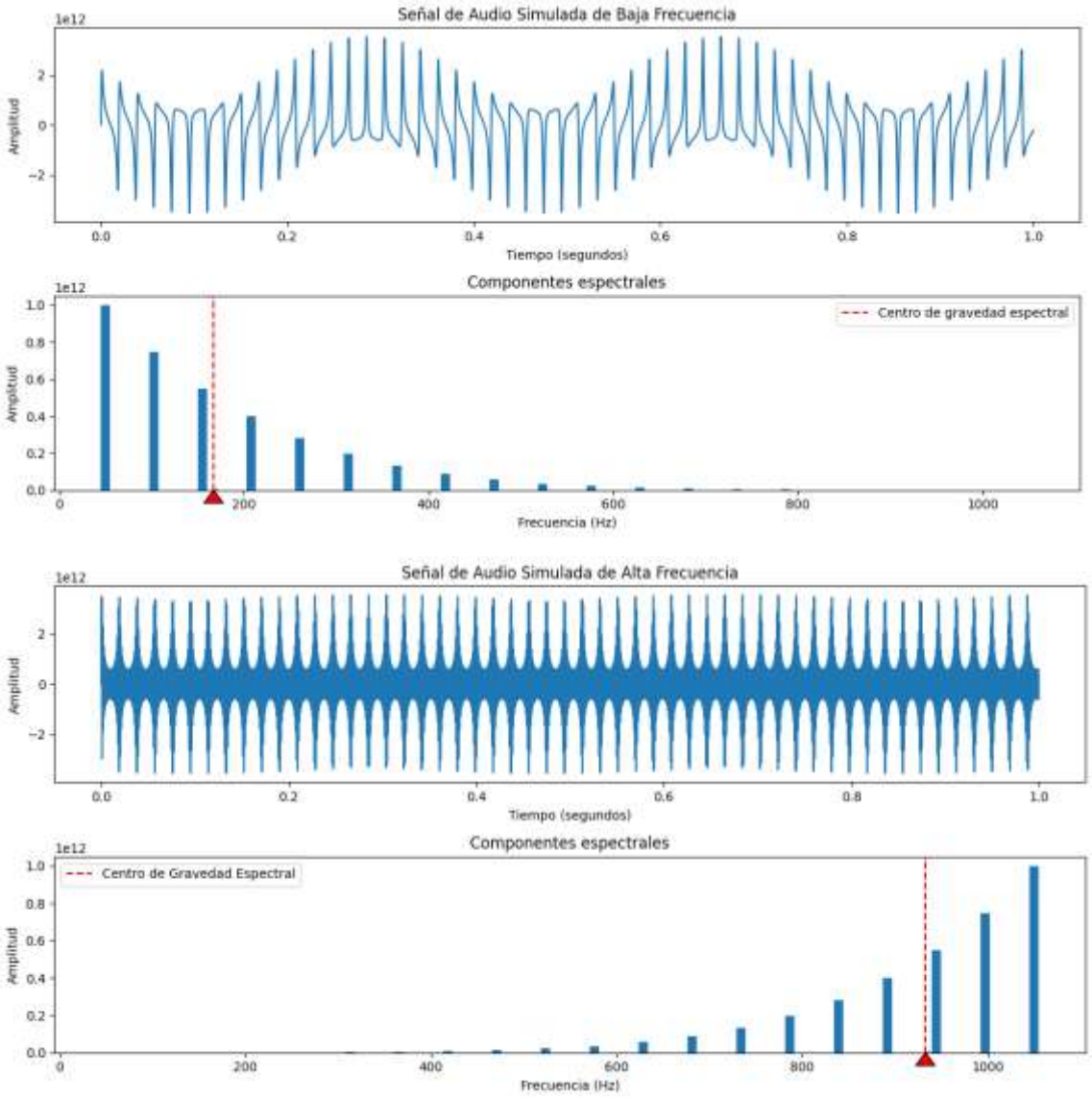


Figura 1. Ejemplos de obtención del centro de gravedad espectral en dos segmentos de audio.

Formalmente, este centro de gravedad se obtiene para un segmento con la siguiente ecuación:

$$C(i) = \frac{\sum_{k=1}^{K} (k+1)\, F_i(k)}{\sum_{k=1}^{K} F_i(k)}$$

Donde es K número de frecuencias en la transformación de Fourier y cada uno de los coeficientes de cada frecuencia en el i-ésimo segmento se representa por $F_i(k)$.

El resultado de este proceso es que la señal de audio representada se transforma en dos señales o funciones: una $E(i)$, la energía, y otra $C(i)$ con los centros de gravedad espectral. El método de Giannakopoulos (2009) indica que cuando los valores de estas dos funciones sean bajos coincide con un tramo de la señal donde no hay habla humana. Por ello, se requiere determinar un umbral para cada una de estas funciones a partir del cual se considera que el valor es bajo. El autor propuso una manera de obtener estos umbrales a partir de las mismas funciones de energía y de centro de gravedad espectral. El método es el siguiente:

1. Construir un histograma de frecuencias con la función de energía $E(i)$
2. Identificar M_1 como el valor de la energía para la barra más alta del histograma.
3. Identificar M_2 como el valor de la energía para la segunda barra más alta del histograma.
4. Obtener el umbral T promediando M_1 y M_2 con la siguiente ecuación $T = \frac{W \cdot M_1 + M_2}{W+1}$. Este promedio está controlado por el parámetro W, el cual cuando $W = 1$ genera un promedio aritmético entre M_1 y M_2 y a medida que W aumenta, en el promedio se pondera más M_1 que M_2. En los experimentos de Giannakopoulos se encontró que $W = 5$ era un valor adecuado para el parámetro, el cual también se usó en este estudio.

En la sección 4.3. en la Figura 2 se ilustra este método mostrando de manera paralela en el tiempo la función de energía (negro), la función de centro de gravedad espectral (cian) y la señal en color negro. En esa figura los umbrales obtenidos se representan con una línea horizontal. Finalmente, en la señal se marca con color rojo los tramos de la señal donde las dos funciones fueron menores que cada uno de sus respectivos umbrales.

3. Antecedentes de investigación

Al estudiar la producción y función de las pausas en distintos discursos comunicativos es posible encontrar diferentes perspectivas de estudio, como por ejemplo, el trabajo de Dall, R., *et al* (2015) quien asegura que cuando se encuentran pausas llenas en un enunciado (ej. “uh” o “um” en inglés), se puede mejorar

significativamente la percepción de naturalidad en un discurso aun cuando se trata de sistemas de síntesis de voz, de modo que no se trata de una deficiencia del lenguaje, sino que puede ser considerada como un rasgo humanizante en procesos modelados de forma artificial.

Con el fin de desarrollar esa investigación, fue utilizado un corpus escrito de noticias con más de 20.000 palabras, así como el corpus recopilado de conversación espontánea recogida en reuniones con 100 horas transcritas a través de 60 oraciones entre ambos corpus. Este trabajo adquiere relevancia en la medida en que, a través de un modelo predictivo, se lograron insertar pausas en el texto para ser posteriormente sintetizadas, lo cual sugiere que existen patrones capaces de predecir el lugar más indicado para ubicar una pausa y conferir de esta manera naturalidad al enunciado. Sin embargo, la metodología utilizada no permite conocer con exactitud cómo se encontraron dichos patrones o cuáles son debido a que se trata de un modelo de caja negra.

No obstante, los distintos usos que se hacen de las pausas pueden exponer a su vez diferencias dialectales de los hablantes, además de estar influenciadas por factores sociales como la educación, la edad y el género de los hablantes. De hecho, la investigación de Pálvölgyi (2020), quien toma 200 enunciados producidos por 16 hombres y 16 mujeres a través de las actividades en el mapa de tareas junto con entrevistas espontáneas subidas a YouTube, conformando así un corpus de 4 horas 29 minutos de audio; permitió concluir que la comprensión de los enunciados por parte del interlocutor no guarda relación con la duración de las pausas. Además, señaló diferencias entre los hablantes del norte y los hablantes del sur de España, siendo estos últimos quienes realizan elisión de fonemas que puede incidir tanto en la duración de las pausas como en la percepción de los hablantes y su comprensión de los enunciados.

En la investigación realizada por Tottie (2017), se incluyen textos de diferentes géneros tales como correos electrónicos, mensajes de chat, mensajes de redes sociales y comentarios en línea recopilando cerca de 870.000 palabras de escritura en inglés norteamericano. Asimismo, se hace uso de un programa de análisis lingüístico llamado CLAN (Computerized Language Analysis), puesto que es necesario identificar y contar el número de instancias de pausas (escritas) en el corpus de datos de manera más precisa con un conjunto de datos extenso. Sin embargo, también se hace uso de técnicas estadísticas que permiten analizar los patrones

de uso de las interjecciones y evaluar si su uso tiene variaciones en función de algunos factores como el género, la edad y el nivel educativo de los hablantes. Gracias al trabajo realizado, fue posible concluir que estas cláusulas son más frecuentes en medio de la oración, antecediendo a frases con nombre propio junto a los adjetivos. Asimismo, se destacan las funciones pragmáticas con las que aparecen en el discurso, las cuales se resumen por un lado en manifestar la postura del autor en el discurso escrito; y por el otro, en tomar un tiempo adicional para poder pensar el siguiente discurso.

Vale la pena destacar trabajos como el de Borzi *et. al* (2017), en el que se analizaron 25 emisiones de habla natural producidas por seis hablantes a través del formato de entrevista con cuatro entrevistados y dos entrevistadores en un total de 10 horas de grabación y veinte minutos recogidos en un corpus de trece entrevistas. Desde la perspectiva de los autores, son los hablantes quienes utilizan distintas estrategias o pistas suprasegmentales como lo son los acentos, el ritmo, la entonación y las junturas, es decir las pausas, para resaltar la información que considera más relevante a su interlocutor (Borzi, *et al*, 2017, p.222).

En resumen, tanto las pausas vacías (silenciosas), o fisiológicas, como las llenas tienen gran relevancia, puesto que dan información sobre el proceso mental que está llevando a cabo el hablante para dar continuidad a su discurso, así como sobre la intención comunicativa que desea lograr, debido a que estas estrategias permiten enfatizar la relación de consecuencia sin necesidad de darle mayor relevancia fonológica a la pronunciación del conector.

De otro lado, existen trabajos que se enfocan en estudiar los rasgos de las pausas en diferentes lenguas con el objetivo de compararlas entre sí como es el caso de la investigación realizada por Crible *et. al* (2017) quienes se proponen estudiar la fluidez del francés y el inglés por medio de un corpus con 10 horas de habla y 100.000 palabras recogidas en llamadas telefónicas, conversaciones, clases, comentarios de deportes, discursos políticos, entre otros. Esta investigación permitió encontrar que las pausas cumplen con funciones diferentes, por lo menos en inglés y francés, dependiendo de si se encuentran agrupadas con marcadores discursivos o no. Por ejemplo, en francés se determinó que ciertas especificaciones en el enunciado surgen específicamente en contextos donde se encuentran tanto marcadores discursivos como pausas, a diferencia del inglés donde las pausas pierden rasgos distintivos en presencia de marcadores discursivos.

En términos generales, los estudios sobre las pausas han usado datos entre dos y treinta y dos hablantes por medio de corpus conformados por grabaciones entre una hora y 10 horas de grabación con el fin de determinar las características acústicas y fonéticas de las pausas, así como la función que tienen en el discurso. Sin embargo, al revisar la metodología utilizada se encuentra con trabajos con corpus que podrían ser más extensos y que dependen en gran medida de las anotaciones del investigador, o de modelos computacionales que no permiten estudiar con detalle el fenómeno lingüístico que explica la aparición de las pausas en ciertos lugares específicos. En ese orden de ideas, las anotaciones manuales realizadas por el investigador y la reducida cantidad de información dificultan el estudio de los factores lingüísticos presentes en el uso de las pausas.

4. Metodología

Para poder analizar los factores lingüísticos presentes en el uso de las pausas silenciosas en audios de entrevistas realizadas a celebridades hablantes de español, es necesario realizar un análisis estadístico que permita identificar de manera sistemática los elementos más frecuentes ante la presencia de pausas y sus duraciones en el discurso. En ese orden de ideas, se requiere de un conjunto de datos amplio que represente el habla espontánea en escenarios de discurso informal conversacional.

En primera instancia, se deben identificar y caracterizar las pausas silenciosas en los audios de los videos de YouTube. Los procesos de identificación y caracterización manual son dispendiosos, lentos y costosos, los cuales serían limitantes para analizar un conjunto de datos grande. Por esta razón, se opta por el uso de métodos automáticos para la identificación y caracterización de las pausas silenciosas, los cuales provienen de los campos del análisis automático de señales y de la inteligencia artificial. En las siguientes subsecciones se describen los datos y métodos automáticos utilizados tanto para la identificación como para la caracterización de las pausas silenciosas en el discurso informal del español.

4.1 Corpus

Para analizar los patrones léxicos alrededor de las pausas producidas entre hispano parlantes, se recolectó un corpus de entrevistas realizadas a 60 celebridades cuya lengua materna fuera el español, tratando de obtener información de la mayoría de países hispanoparlantes a través de cantantes, políticos, actores, deportistas y escritores provenientes de 12 países, tal como se puede apreciar en la tabla 1, la cual contiene el detalle de celebridades por país y dominio seleccionados en el corpus.

País	Celebridades	Videos	Horas
Argentina	2 deportistas	119	24,8
Bolivia	1 político	221	106,7
Colombia	8 cantantes	382	77,6
	1 político	7	7,27
	1 actor	3	0,2
	3 deportistas	187	20,2
Chile	1 actor	34	2,4
Cuba	2 cantantes	58	6,9
Estados Unidos	1 cantante	10	1,7
El Salvador	1 político	3	3
España	3 cantantes	32	6,8
	2 actores	21	3,5
Nicaragua	1 cantante	83	13,3
Panamá	1 cantante	7	7,7
Perú	1 cantante	11	5,2
	1 escritor	57	23,2
Puerto Rico	3 cantantes	180	23
Uruguay	1 cantante	2	1,5
Venezuela	3 cantantes	25	7,8
	1 político	5	

Tabla 1. Distribución de las celebridades del corpus por país y dominio

Las entrevistas fueron obtenidas a través de videos disponibles en YouTube, por lo que se usó la url https://www.youtube.com/results?search_query= y se adicionó a la búsqueda el nombre de la celebridad con la palabra “Entrevista”. Los videos resultantes se agregaron, a excepción de aquellos videos

con publicidad. No obstante, al seguir esta metodología la cantidad de videos puede ser demasiado extensa puesto que siguen apareciendo videos relacionados a medida que se baja en el buscador.

Por esta razón, fue necesario aplicar un criterio que permita detener la recolección de videos en la lista, teniendo en cuenta que los primeros videos son aquellos con una relevancia mayor en la búsqueda, mientras que los videos obtenidos al aumentar la lista tienen mayor posibilidad de ser poco relevantes. El criterio utilizado fue encontrar 7 videos sin el nombre de la celebridad en el título, puesto que así se consigue una buena cantidad de videos pertinentes para la investigación, garantizando un equilibrio en el corpus sin omitir demasiados datos de interés, y a la vez, sin agregar videos innecesarios. El valor de este parámetro se estableció por observación, a través de ensayo y error.

Al aplicar este método se obtuvo un total de 3.002 videos, los cuales fueron posteriormente seleccionados cuidadosamente para evitar contar con videos que no fueran entrevistas, que estuvieran en una lengua distinta del español; y finalmente, que no fueran entrevistas a celebridades; luego de este proceso manual, el corpus se redujo a 1.460 videos en el corpus con un total de 347 horas de entrevistas. Este proceso de depuración manual tardó aproximadamente 60 horas y fue realizado por el autor.

4.2 Transcripción Automática con Marcas de Tiempo

El primer proceso aplicado al corpus, con el objetivo de identificar las pausas silenciosas, consistió en aplicar Subtitulado Sincronizado Automático utilizando la herramienta Whisper-timestamped (ver subsección 2.3). Para esto, se obtuvieron los audios asociados a cada video en formato .mp4 con la herramienta *pytube*[1]. Luego se utilizó Whisper-timestamped , de modo que se consiguió para cada video una lista de tuplas que contienen la transcripción de la palabra detectada, la marca de tiempo de su inicio y la marca de su terminación. Las marcas de tiempo se obtienen en segundos con una resolución de 10 milisegundos. De otro lado, los audios de los videos recolectados en el corpus (ver subsección 4.1.) fueron procesados usando el modelo “tiny” de Whisper utilizando un entorno de programación basado en una CPU, de manera que cada

[1] https://pytube.io/en/latest/

video se transcribe y se marcan los tiempos de cada palabra a una velocidad comparable con la duración del video. Este proceso de transcripción duró aproximadamente 350 horas (15 días aproximadamente) en un Notebook de Google Collaboratory. En total se detectaron 2.995.580 palabras en el corpus con un vocabulario de 132.037 palabras. Hasta donde se tiene conocimiento, es la primera vez que se usa la tecnología Subtitulado Sincronizado Automático para investigación lingüística a nivel fonético y en especial para el análisis de pausas.

Debido a que el propósito del Subtitulado Sincronizado es la presentación de las palabras en un video, las pausas entre las palabras no son parte de su objetivo. Por esta razón, cuando en el discurso analizado hay una secuencia hablada de palabras, las marcas de tiempo de terminación de una palabra coinciden con la marca de tiempo de inicio de la siguiente. Esto hace que las pausas en general de menos de 1 segundo no sean identificadas por la herramienta.

4.3 Identificación Automática de Pausas Silenciosas

Frente a la limitación de no poder identificar las pausas de manera fiable en la subsección anterior, se utilizó el método de Giannakopoulos (2009) descrito en la subsección 2.4. programado a través de un Cuaderno de Google Colab. Se usaron los audios en formato .mp4 obtenidos en la subsección anterior. Estos audios fueron decodificados con la herramienta *librosa*[2] la cual obtiene de cada archivo de audio una lista numérica con la digitalización de la señal de audio en una rata de muestreo de 22.050Hz. El método de Gianakopoulos se implementó usando un tamaño de ventana de análisis, o segmento, de 50ms (sugerido por el mismo autor) y se analizaron segmentos de esa longitud cada 5ms. El método original analizaba segmentos cada 50ms, lo que significa que no había sobrelapos entre dos segmentos contiguos. Sin embargo, esta decisión probablemente estuviese asociada a limitaciones en capacidad de cómputo, las cuales no existen en la actualidad. En ese sentido, para mejorar la resolución de las funciones de energía y centro de gravedad espectral, se analizaron los segmentos cada 5ms produciendo un sobrelapo de 45ms en segmentos contiguos. Para el análisis espectral, o sea la Transformada Discreta de Fourier se utilizó la herramienta *scipy.signal*[3]. Para el cálculo automático de los umbrales se utilizaron histogramas de 20 intervalos (barras) con un

[2] https://librosa.org/doc/latest/index.html
[3] https://docs.scipy.org/doc/scipy/reference/signal.html

parámetro $W = 5$ (sugerido por Giannakopoulos). La cantidad de 20 intervalos se consideró adecuada valorando visualmente pocas irregularidades y transiciones suaves en las frecuencias de los intervalos.

En la Figura 2 se visualiza una muestra de una señal de audio con sus funciones de energía (negro) y centros de gravedad espectral (cian). De igual modo, en los mismos colores correspondientes se ilustran los umbrales detectados automáticamente para cada función. Se resaltan en rojo los dos segmentos de la señal donde tanto la función de energía como la de los centros de gravedad espectral están por debajo de sus respectivos umbrales. Conviene señalar que en esa muestra de audio aparentemente hay un silencio en el centro de la gráfica, el cual es descartado por el método ya que la función de centro de gravedad espectral no lo identifica como un tramo de audio de frecuencias bajas. Por esta razón, no es considerado "ruido" de fondo sino habla humana. En contraste, los dos segmentos marcados en rojo claramente se descomponen en frecuencias bajas. En este punto se identifica para el audio de cada video todos los segmentos "silenciosos" o con ruido de fondo que no tienen presente habla humana.

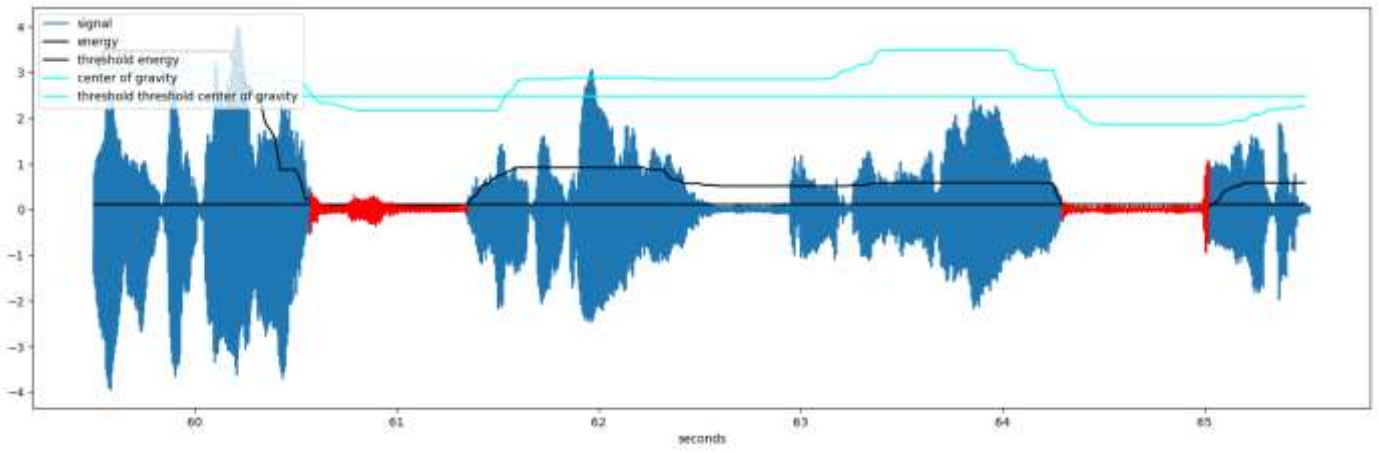


Figura 2. Señal de audio, energía, centro de gravedad espectral y umbrales (líneas horizontales) de un segmento del video en https://www.youtube.com/watch?v=6vaR7APyJrE

Ahora se combina la información de las anotaciones de Whisper-timestamp, en la cual están identificados aproximadamente los tiempos donde son pronunciadas las palabras, con la lista de los tramos sin actividad

de voz identificados en la señal de audio. Sumando estas dos fuentes de información, se identifican cuáles son las palabras de los diálogos entre las cuales hay pausas silenciosas. En la Figura 3 se ilustra un ejemplo de una señal de audio que contiene habla con las marcas de tiempo identificadas por Whisper-timestamped y los tramos sin actividad de habla detectados con el método de Giannakopoulos (en rojo). En la figura se aprecia que el tramo identificado a la izquierda tanto el método de Giannakopoulos como Whisper-timestamped coinciden con la identificación de una pausa. Sin embargo, en el tramo del lado derecho el método de Giannakopoulos identificó un segmento sin habla humana, pero las anotaciones de Whisper-timestamped son continuas.

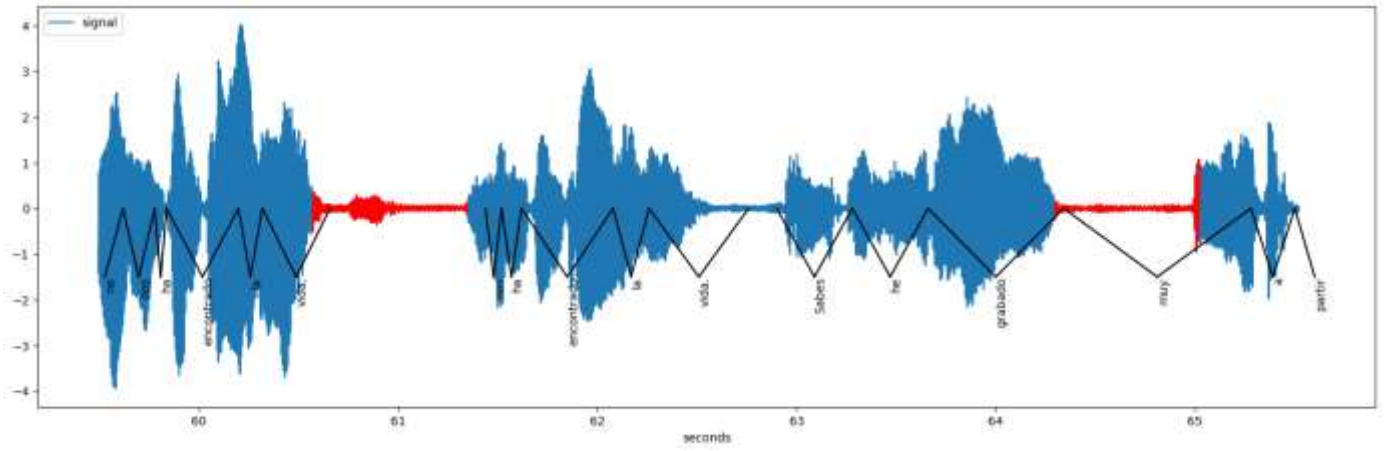


Figura 3. Muestra de audio del video https://www.youtube.com/watch?v=6vaR7APyJrE con anotaciones de Whisper-timestamped y segmentos sin habla humana identificados con el método de Giannakopoulos.

Es importante señalar que la duración de esas pausas está determinada por las duraciones de los tramos sin actividad de voz y no por las marcas de tiempo obtenidas con Whisper-timestamped. Esto debido a que como se muestra en la Figura B el método de Giannakopoulos provee información temporal precisa de la identificación de los tramos silenciosos, los cuales o son ignorados, o son identificados de manera imprecisa por Whisper-timestamped. Para realizar el proceso de detección de las pausas silenciosas con la información disponible, se identificaron varios casos de interacción entre las marcas de tiempo de las palabras y los silencios de la señal, los cuales se presentan a continuación.

Caso #1. Cuando Whisper-timestamped ha identificado una pausa entre sus anotaciones y en ese mismo segmento el método de Giannakopoulos (2019) detecta un segmento sin habla humana. En este caso se tiene una coincidencia que permite identificar con confidencia una pausa silenciosa entre dos palabras. La Figura 4 muestra un ejemplo de este caso. De manera empírica se consideran solo las pausas silenciosas donde el segmento sin habla humana tiene una duración de al menos el 70% de la pausa identificada con Whisper-timestamped. Lo anterior para descartar segmentos de audio sin habla humana demasiado cortos. Adicionalmente, en la figura se aprecia que la duración del segmento marcado en rojo representa mejor la duración de la pausa que las marcas de tiempo de Whisper-timestamped.

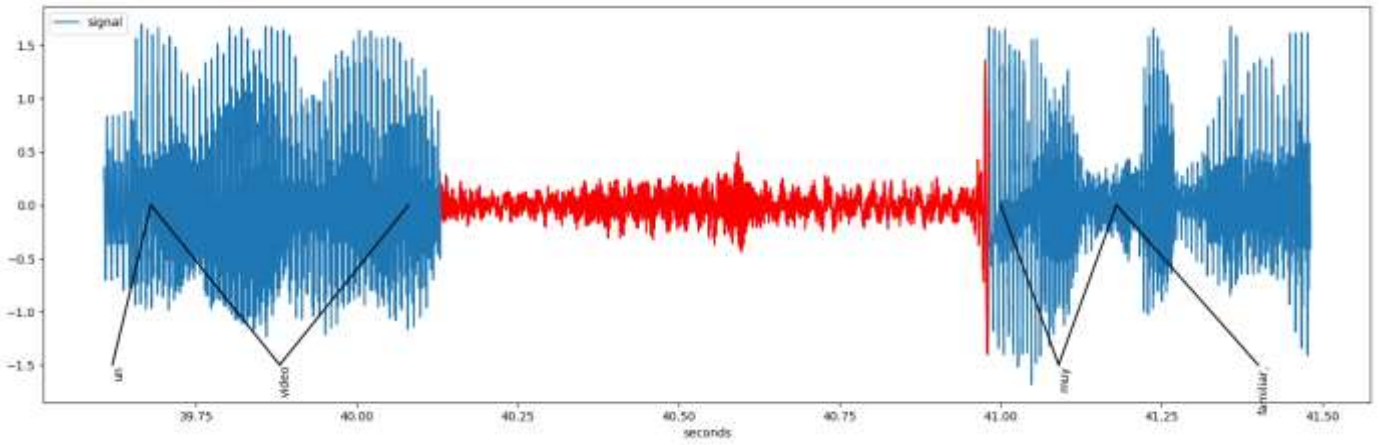


Figura 4. Muestra del video .https://www.youtube.com/watch?v=lmiAfgGFC2Q ilustrando un silencio de Caso #1

Caso #2. A diferencia del Caso #1 donde la anotación de Whisper-timpestamped inicia antes del segmento sin habla humana, y acaba después del mismo, este caso se presenta cuando las marcas de tiempo de Whisper-timestamped coinciden con segmentos sin habla humana (en rojo) tanto en el principio de la pausa como en la finalización, es decir, que la anotación se encuentra dentro del segmento sin habla, evidenciando de este modo una pausa silenciosa. La Figura 5.1 ilustra un ejemplo donde el segmento sin habla humana es de mayor duración que la pausa de Whisper-timestamped. Por otro lado, la Figura 5.2 presenta otro caso interesante donde Whisper-timestamped no identificó pausa aunque la transición entre dos palabras coincide con un segmento sin habla humana, revelando así la pausa silenciosa entre las palabras “identifica” y “ese”.

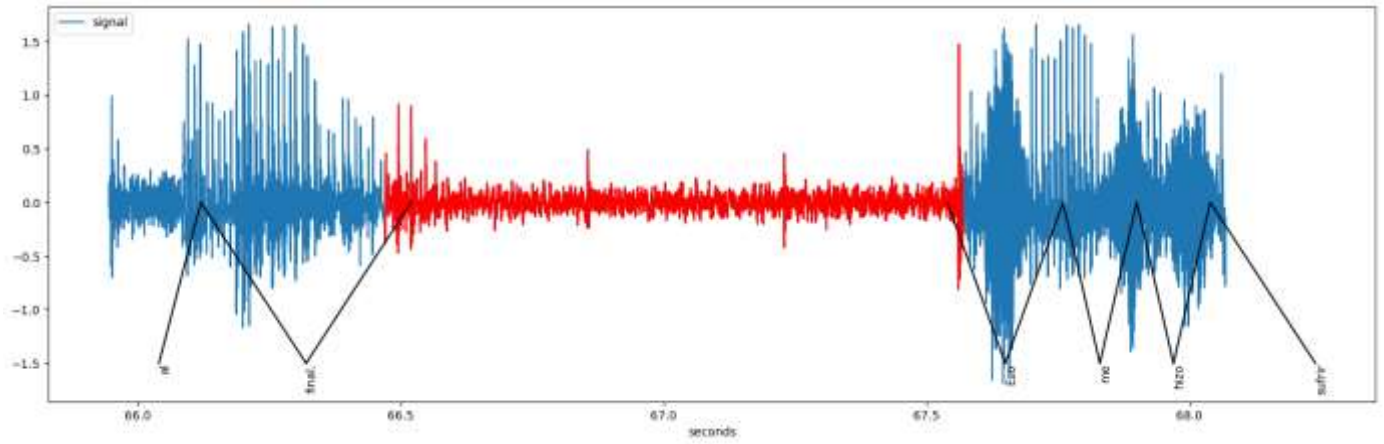


Figura 5.1. Muestra del video://www.youtube.com/watch?v=ykc_D2aIdYI **con un silencio de Caso #2**

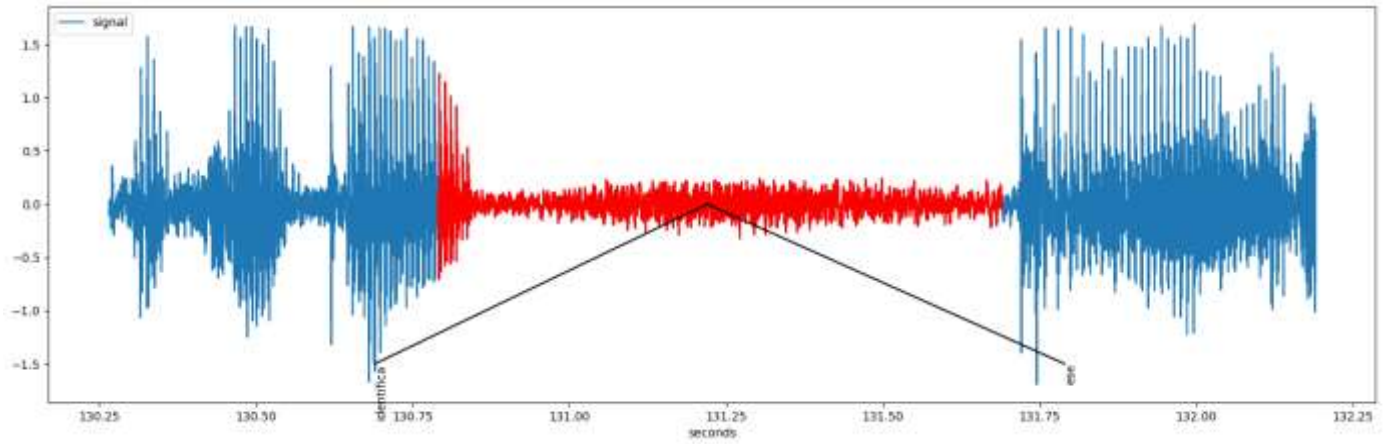


Figura 5.2. Muestra del video://www.youtube.com/watch?v=ykc_D2aIdYI **con un silencio de Caso #2**

Casos #3 y #4. Se presentan en aquellos casos en que Whisper-timestamped detecta una pausa, mientras que el segmento sin habla coincide solo con una de las marcas de tiempo (la de inicio o la finalización). La Figura 6 muestra un ejemplo del Caso #3 cuando el segmento sin habla coincide con la marca de tiempo de inicio de la pausa, lo que contrasta con el Caso #4 donde el segmento sin habla presenta coincidencia con la marca de tiempo de la finalización de la pausa.

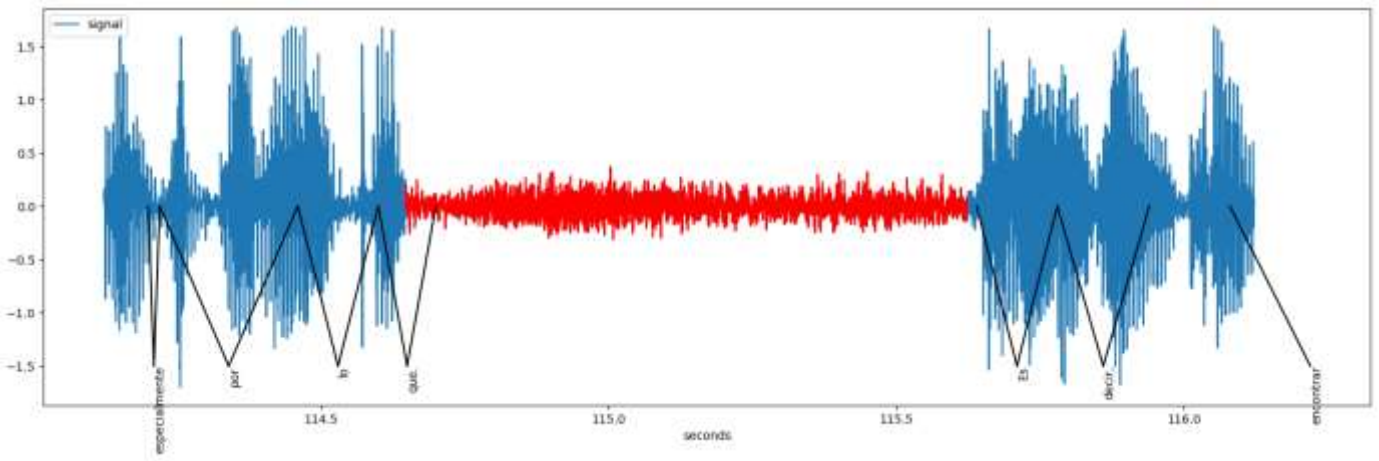


Figura 6. Muestra del video: https://www.youtube.com/watch?v=G6qUXvq8mqM **con un silencio de Caso #3**

Los cuatro casos descritos anteriormente fueron los que identificaban con mayor claridad las pausas silenciosas, puesto que se presentaban otros casos que resultaban ambiguos, como cuando ocurrían dos o más segmentos sin habla humana en una pausa entre las marcas de tiempo; o donde en ciertos segmentos se combinaban pausas entre las marcas de tiempo y los segmentos sin habla, haciendo difícil identificar un patrón reproducible. Resulta evidente que, al usar solamente los cuatro casos presentados, un número indeterminado de pausas fueron ignoradas; no obstante, al realizar un análisis visual de las pausas detectadas, la gran mayoría son pausas identificadas correctamente.

En este punto es preciso decir que, debido a la magnitud del corpus y las limitaciones de tiempo, no fue posible hacer una estimación cuantitativa de la precisión de la detección a través de la detección manual de las pausas. Sin embargo, se espera alcanzar el orden de cientos de miles de pausas detectadas por medio de estos cuatro casos, con el fin de encontrar los patrones estadísticos presentados en los resultados a pesar del ruido generado por las pocas falsas pausas detectadas.

4.4 Filtrado De Pausas Silenciosas

Frente a la duración promedio de una pausa hay distintas posturas. En el trabajo realizado por Grosjean y Deschamps (1975) se encuentra que las pausas silenciosas se distribuyen en su mayoría entre 250 y 374

milisegundos. Blondet por su parte (2006) señala que las pausas presentan una detención prolongada con una duración igual o superior a los 350 milisegundos, mientras que autores como Wang *et. al* (2010) afirman que los valores promedio de duración de un grupo respiratorio (pausa) se encuentran entre 2,42 y 3,84 segundos en habla espontánea. Otros investigadores como Lundholm (2015) entienden que la duración de las pausas se encuentra en un rango entre 550 milisegundos y 3.000 milisegundos. Teniendo en cuenta estos trabajos, se decidió que la pausa debía ser de máximo dos segundos dado que el corpus, como se ha mencionado, tiene presente bastante ruido, así como arreglos de edición para presentar los videos, de manera que el valor de dos segundos permite encontrar mejor la mayor cantidad de pausas realizadas por los hablantes en el contexto comunicativo de entrevistas.

En el corpus seleccionado no es posible diferenciar con claridad los enunciados entre hablantes puesto que únicamente se obtiene la transcripción de lo que se dijo, pero no se marca quién lo dijo. Es posible encontrar pausas que sean producto de un hablante, aunque también es posible que los silencios encontrados sean generados por cambio de turno de los hablantes. Para evitar esta confusión y asegurar que las pausas encontradas son realizadas por el mismo hablante, se decide utilizar las pausas intermedias, es decir, aquellas que ocurren entre dos palabras y donde Whisper no detectó signo de puntuación entre las dos palabras. De igual manera, se deciden incluir las pausas gramaticales, que son aquellas donde sí se transcribió un signo de puntuación, escogiendo la "," puesto que se asume que no hay un cambio de turno, a diferencia del punto "." que sí puede generar ambigüedad sobre si es el mismo hablante o si es un cambio de turno, junto al signo de interrogación (?) que señala una pregunta dirigida al interlocutor para que intervenga en la conversación.

4.5 Sistematización y Análisis de Datos

Gracias al proceso detección automático descrito en la sección anterior se puede obtener una lista de pausas silenciosas con la siguiente información para cada una:

1. Palabra antes de la pausa (cadena de caracteres)
2. Signo de puntuación al final de la palabra antes de la pausa (carácter: ninguno, ',') para distinguir pausas intermedias de gramaticales.
3. Duración de la pausa (segundos con tres decimales)
4. Tiempo inicio de la pausa en el audio (segundos con tres decimales)

5. Palabra después de la pausa (cadena de caracteres)
6. El caso de detección (1, 2, 3, o 4)

La presencia o no del signo de puntuación "," después de la palabra y antes de la pausa es resultado de las anotaciones de Whisper, lo que permite distinguir las pausas que coinciden con una frontera de una frase (pausas gramaticales), de las pausas intermedias.

Los siguientes datos, para cada pausa, fueron ampliados utilizando información del mismo corpus o de fuentes externas:

1. Frecuencia total de la palabra antes de la pausa en el corpus
2. Frecuencia total de la palabra después de la pausa en el corpus
3. Frecuencia del bigrama conformado por las palabras antes-después en el corpus
4. El dominio de la profesión de la celebridad asociado al video (cantante, político, actor, deportista y escritor)
5. Número de sílabas de la palabra antes de la pausa usando la herramienta *silabeador*[4]
6. Número de sílabas de la palabra después de la pausa usando la herramienta *silabeador*
7. Posición de la sílaba tónica en la palabra antes de la pausa usando *silabeador* (aguda, grave o esdrújula)
8. Posición de la sílaba tónica en la palabra después de la pausa usando *silabeador*

En total se dispone de 14 datos para caracterizar cada una de las pausas detectadas. Vale la pena señalar que en cada pausa siempre existen los primeros 10 datos, mientras que los 4 últimos dependen si en la "palabra" la herramienta *silabeador* identificó sílabas válidas del español, puesto que valores como las cantidades y años, identificadas con cifras numéricas, no se cuentan como sílabas.

Además, para cada video se obtuvieron los siguientes datos:

1. Duración del video desde el inicio de la primera palabra detectada hasta el final de la última

[4] https://pypi.org/project/silabeador/1.0.2.post14/

2. El número de palabras detectadas en el video.
3. Rata de producción de las palabras por video haciendo el cociente 2, sobre 1.
4. Duración promedio de las pausas detectadas en el video

Estos datos adicionales a nivel de video se recolectaron con el objetivo de confirmar o refutar la hipótesis de Zellner (1994) quien relaciona la duración de las pausas con la velocidad de producción de palabras en el discurso.

Los resultados en la siguiente sección se obtienen a partir de los datos sistematizados descritos antes y se presentan a través de histogramas, diagramas de cajas (*box-plots*), tablas comparativas y correlaciones. En el caso de estas últimas, se utilizó la correlación de Spearman la cual resulta más conveniente, ya que no asume normalidad en las variables y es una estadística no paramétrica. Esta se calculó usando la herramienta *scipy.stats*[5]

5. Resultados

Los resultados de esta sección buscan caracterizar las pausas silenciosas del español basados en las ocurrencias de estas pausas identificadas con la metodología automática presentada en la sección anterior. Algunos de los resultados que se presentan son similares a los de otros estudios y se presentan a modo de comparación. Otros resultan de carácter novedoso debido a que la naturaleza de este estudio permite encontrar un mayor número de pausas estudiadas en comparación con estudios anteriores, lo que significa que es posible presentar distribuciones de ciertas características que antes no era posible.

Aplicando la metodología de la Sección 4 en el corpus recolectado, se identificaron 197.720 pausas silenciosas ajustando los siguientes parámetros del modelo: duración de la ventana de análisis $v = 50ms$; porcentaje de sobrelapo de las ventanas $s = 90\%$; relacionados con el método de Giannakopoulos (2009) el parámetro de ajuste de umbrales $W = 5$ y el número de intervalos $b = 20$. Dado que el corpus tiene 347 horas de diálogos (20.815 minutos), la rata promedio de pausas detectadas por minuto es de 12,3 pausas por

[5] https://docs.scipy.org/doc/scipy/reference/stats.html

minuto. De cada pausa se dispone de la siguiente información: 1) el caso de tipo de pausa identificada, 2) la palabra antes de la pausa, 3) la palabra después de la pausa, 4) el tiempo en el video en milisegundos de inicio de la pausa, y 5) el tiempo de terminación. Para caracterizar la efectividad de los posibles casos de pausas presentados, en la Figura 7 se muestra en un diagrama de cajas (*box plot*) la distribución de las pausas identificadas con respecto a los casos presentados en la subsección 4.3. Se observa que el 85% de las pausas corresponden a los casos #2 y #3, y que las pausas más cortas corresponden a los casos #1 y #4.

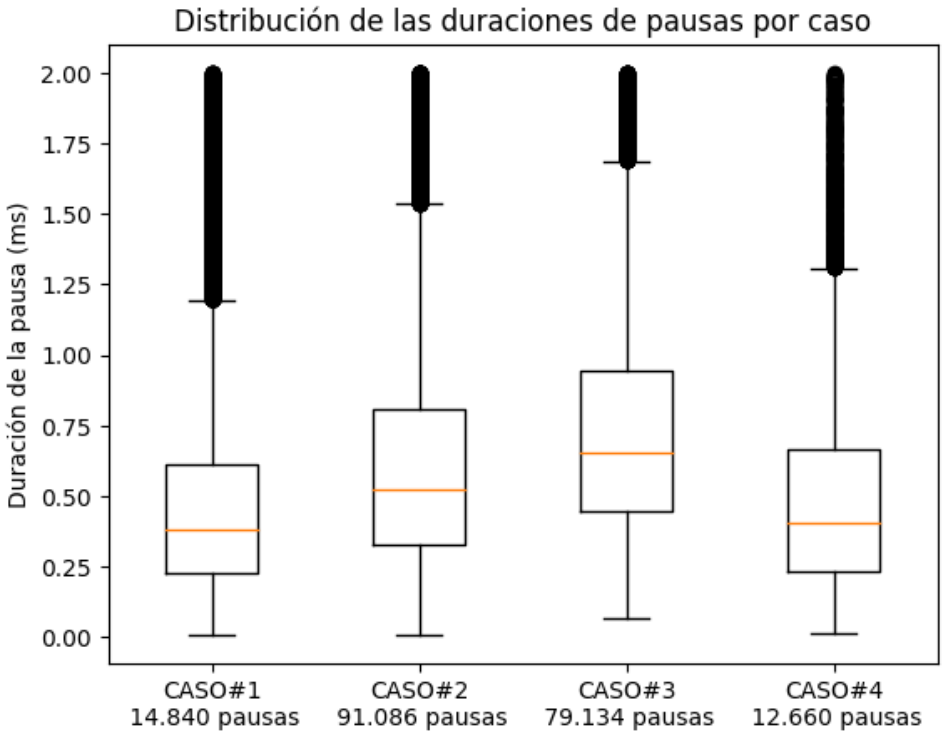


Figura 7. Distribución de las pausas silenciosas detectadas por caso.

En la Tabla 2 se presentan los rasgos léxicos más frecuentes del corpus y de las pausas silenciosas identificadas. Estos se obtuvieron recorriendo cada una de las palabras del corpus actualizando conteos en los tiempos donde se identificaron automáticamente las pausas silenciosas. Gracias a que el número de palabras en el corpus, junto al número de pausas, es considerable; es posible identificar patrones en las frecuencias. En la tabla se reportan los 10 elementos más frecuentes de las palabras de corpus, así como las palabras antes de las pausas, los bigramas palabra antes-[Pausa]-palabra después de las pausas y de las palabras después de las pausas. El objetivo es comparar si existe una relación entre las frecuencias de las

palabras alrededor de las pausas con las palabras frecuentes del corpus. Vale la pena destacar que las frecuencias de las palabras que se encuentran después de las pausas son considerablemente mayores que las palabras antes de las pausas.

Ranking	Palabra	frecuencia	Palabra antes	frecuencia	Bigrama antes-después	frecuencia	Palabra después	frecuencia
1	que	133.712	que	3.679	no [P] no	361	y	19.068
2	de	127.039	de	2.091	que [P] es	248	que	12.286
3	la	92.629	no	1.736	sí [P] sí	200	de	6.877
4	no	91.052	es	1.545	que [P] se	176	en	5.942
5	y	86.752	y	1.487	años [P] y	169	no	5.743
6	a	76.209	la	1.447	es [P] que	136	pero	5.409
7	el	66.707	entonces	1.201	lo [P] que	134	es	4.618
8	en	66.442	eso	1.192	que [P] no	125	a	4.493
9	es	57.400	más	1.166	bien [P] y	120	el	4.269
10	un	40.080	años	1.093	vida [P] y	110	la	3.662
11	lo	34.554	el	984	en [P] el	109	yo	3.151
12	yo	31.091	sí	959	cosas [P] que	109	con	2.976
13	se	31.067	en	920	eso [P] y	107	para	2.816
14	me	30.687	bueno	866	país [P] y	106	porque	2.776
15	con	29.609	yo	830	que [P] el	105	por	2.594
16	los	28.547	los	828	que [P] que	103	se	2.593
17	por	27.146	a	808	que [P] la	102	como	2.545
18	una	25.364	bien	804	no [P] es	93	un	1.951
19	pero	22.908	también	796	que [P] yo	89	si	1.940
20	como	22.349	vida	727	gente [P] que	87	o	1.673

Tabla 2. Palabras más frecuentes: i) en el corpus, ii) antes de las pausas silenciosas, iii) bigrama antes-[Pausa]-después, y iv) palabras después de la pausa.

La Figura 8 muestra la distribución de la duración de las pausas identificadas en el corpus. Al igual que en la Figura 7, en el eje vertical se codifica la frecuencia y en el horizontal el tiempo de duración de las pausas

medido en segundos. Se aprecia que la mayoría de las pausas son de alrededor de 250 ms y que su frecuencia disminuye rápidamente hasta el valor de corte típico de 2 segundos para ser considerada como una pausa en el habla.

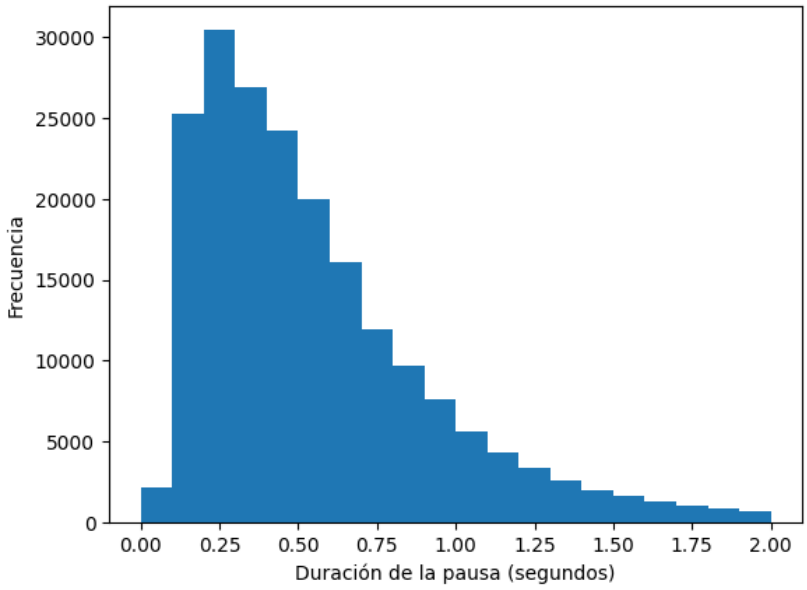


Figura 8. Distribución de la duración de las pausas en el corpus.

En la Figura 9 se muestra la distribución de sílabas de las palabras antes y después de las pausas. La silabación se obtuvo usando el silabeador para español para Python[6]. En el eje vertical se muestra el conteo de palabras y en eje horizontal el número de sílabas. La figura ilustra las distribuciones de la cantidad de sílabas de las palabras antes (azul) y después (rojo) las cuales se sobreponen. En cada columna se clarifica el número de sílabas correspondiente. Se evidencia que las palabras antes de las pausas son más largas, mientras que las palabras después de la pausa suelen tener menos sílabas y consecuentemente ser más cortas.

[6] https://pypi.org/project/silabeador/1.0.2.post14/

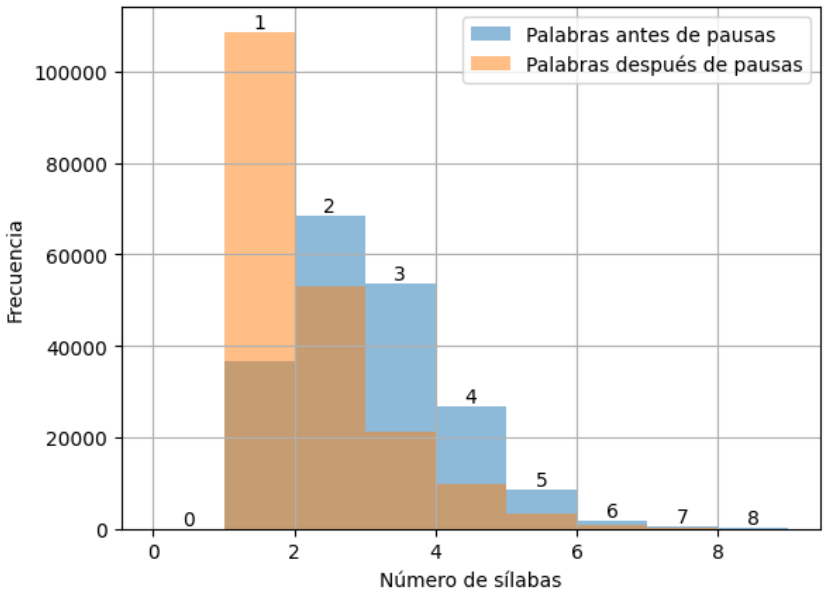


Figura 9. Distribución del número de sílabas antes y después de la pausa.

La Figura 10 presenta la distribución de las pausas de acuerdo con la función gramatical que cumplen las pausas intermedias. Esta información se obtiene gracias a las anotaciones obtenidas con Whisper en las cuales se hace marcación de la puntuación, lo cual permite distinguir las pausas donde aparecen marcas gramaticales de aquellas en las que no. Debido a que se están analizando las pausas silenciosas intermedias en el discurso de un hablante y no en los cambios de turno de interlocutor, las pausas gramaticales consideradas son solamente las que coinciden con la marcación de la coma. Se busca comparar si hay diferencias en las duraciones entre las pausas intermedias y las que se encuentran en una frontera gramatical. Al realizar la comparación, se puede apreciar que las pausas intermedias son más cortas que las gramaticales.

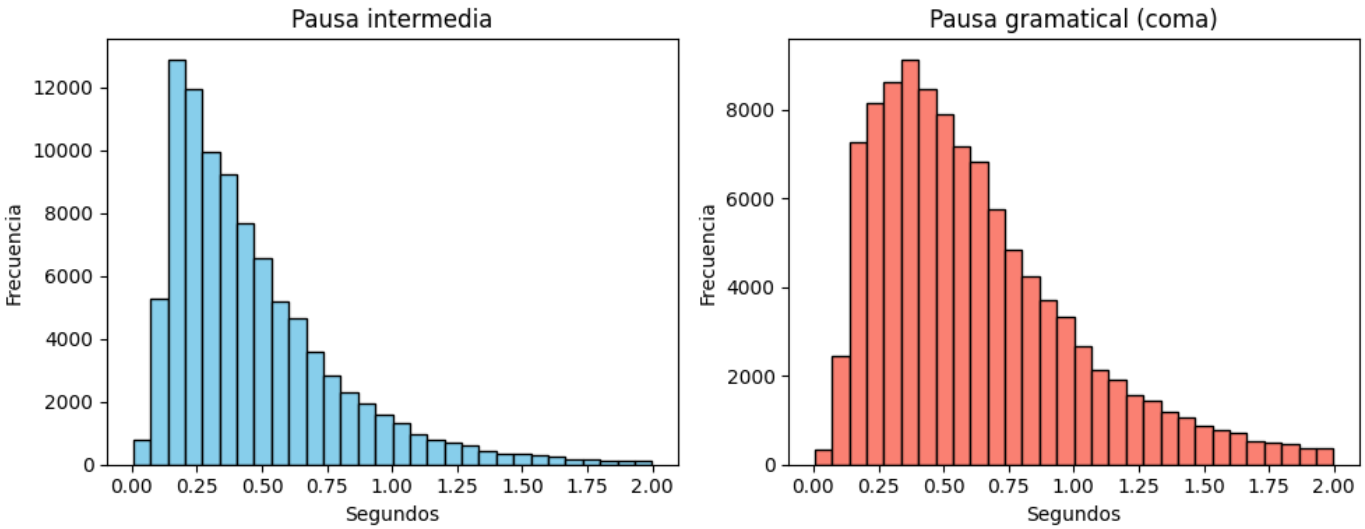


Figura 10. Distribución de la duración de las pausas intermedias y gramaticales

Dado que el español es una lengua con acentuación gráfica, es posible observar el número de pausas encontradas según el acento de las palabras antes y después de las mismas. La Tabla 3 muestra los conteos de las transiciones de la acentuación de la palabra antes de la pausa a la acentuación de la de después en pausas intermedias y gramaticales. En ese orden de ideas, es evidente que el patrón más frecuente es el de palabras graves antes de la pausa con palabras agudas después de la pausa, el cual tiene la mayor frecuencia de 77.222 (aproximadamente del 40% de las pausas), seguido por la palabra grave antes y después de la pausa con un total de 43.119 apariciones y la palabra aguda antes y después de la pausa con un total de 40.011. Como parámetro de comparación se presenta la distribución de acentos en la totalidad de las palabras del corpus, que es de 1.724.114 palabras agudas, 1.208.375 palabras graves y 48.375 palabras esdrújulas.

Tónica antes [PAUSA] Tónica después	**Todas**	**Gramaticales**	**Intermedias**
grave [PAUSA] aguda	77.222	43.548	33.674
grave [PAUSA] grave	43.119	27.271	15.848

aguda [PAUSA] aguda	40.011	17.718	22.293
aguda [PAUSA] grave	26.387	9.735	16.652
esdrújula [PAUSA] aguda	4.359	2.648	1.711
esdrújula [PAUSA] grave	2.454	1.725	729
grave [PAUSA] esdrújula	1.089	456	633
aguda [PAUSA] esdrújula	973	175	798
esdrújula [PAUSA] esdrújula	125	74	51
TOTALES	195.739	103.350	92.389

Tabla 3. Distribución de las pausas según el acento de la palabra antes y después de la pausa, y según si el tipo de pausa es gramatical o intermedia.

La Tabla 4 presenta las correlaciones encontradas entre diferentes variables presentes en el corpus asociadas a cada pausa, que son la frecuencia de las palabras antes y después de la pausa, el número de letras que tienen las palabras antes y después de la pausa junto con la duración de la pausa. El principal objetivo es estudiar si existe alguna correlación entre estas variables, especialmente la relación entre la frecuencia de aparición de la palabra y la duración de la pausa, debido a que desde la postura de Brysbaert (2011), las palabras se recuperan y producen con mayor o menor eficiencia (tiempo) según el número de veces que aparezcan en el discurso (también conocido como *word frequency effect*), lo cual podría afectar la duración de las pausas. Dentro de las correlaciones más altas se encuentra la frecuencia del número de letras de la palabra antes de la pausa con la duración con un valor de 0,172 mientras que la frecuencia de palabras antes de la pausa guarda una correlación negativa del -0,165 con la duración de la pausa. Los valores de p de las correlaciones fueron todos muy cercanos a cero debido a la gran cantidad de pausas comparadas $n = 197.720$.

Variable	Correlación con la duración de la pausa
Frecuencia de palabra antes de pausa en el corpus	**-0.165**
Frecuencia de palabra después de pausa en el corpus	0.109
Número de letras palabra antes de pausa	**0.172**
Número de letras palabra después de pausa	-0.100

Tabla 4. Correlaciones entre la frecuencia de palabras, número de letras versus la duración de las pausas.

En cuanto a la correlación entre la frecuencia de transiciones y duración de pausa se calcularon las frecuencias de todos los bigramas del corpus, es decir las posibles combinaciones de palabras antes y después sin tener en consideración el factor de pausa. Luego, se obtuvo el bigrama en presencia de pausa para evaluar si hay relación entre la frecuencia de transición de palabras y la duración de la pausa. En otras palabras, evaluar si cuando la combinación de palabras es más frecuente tiene una relación con la duración de la pausa. Se encontró una correlación inversa del $\rho = -0{,}097$. Entre más alta es la frecuencia de transición más corta es la duración de la pausa.

Por otra parte, se calculó la correlación entre la duración promedio de las pausas de los videos y la rata de palabras por segundo de cada video para confirmar la hipótesis de Zellner (1994) sobre la duración de las pausas. Esta autora afirma que la duración de la pausa no depende de un factor fisiológico independiente como la respiración, sino que esta se acomoda al ritmo en el que están hablando. La correlación entre estas dos variables fue de $\rho = -0.416$ con un valor p muy cercano a cero. Esto indica que, entre más rápido se dicen las palabras, es menor la duración de las pausas.

La Figura 11 presenta la distribución de la duración de pausas por dominio con el objetivo de determinar si según el contexto comunicativo la duración de la pausa es mayor o menor. Para realizarlo, las pausas fueron agrupadas en cinco dominios: actores, cantantes, políticos, deportistas y escritores. Esta clasificación se hizo asignando manualmente el dominio a cada celebridad, luego esta clasificación se propagó hacia los videos de cada celebridad y finalmente se trasladó a las pausas identificadas. Posteriormente, fueron agrupadas entre sí y representadas gráficamente a través del gráfico de cajas (*box plot*). Se observa que la duración más alta es la del dominio de políticos, seguida por el dominio de escritores, mientras que las duraciones más bajas pertenecen al dominio de cantantes y actores.

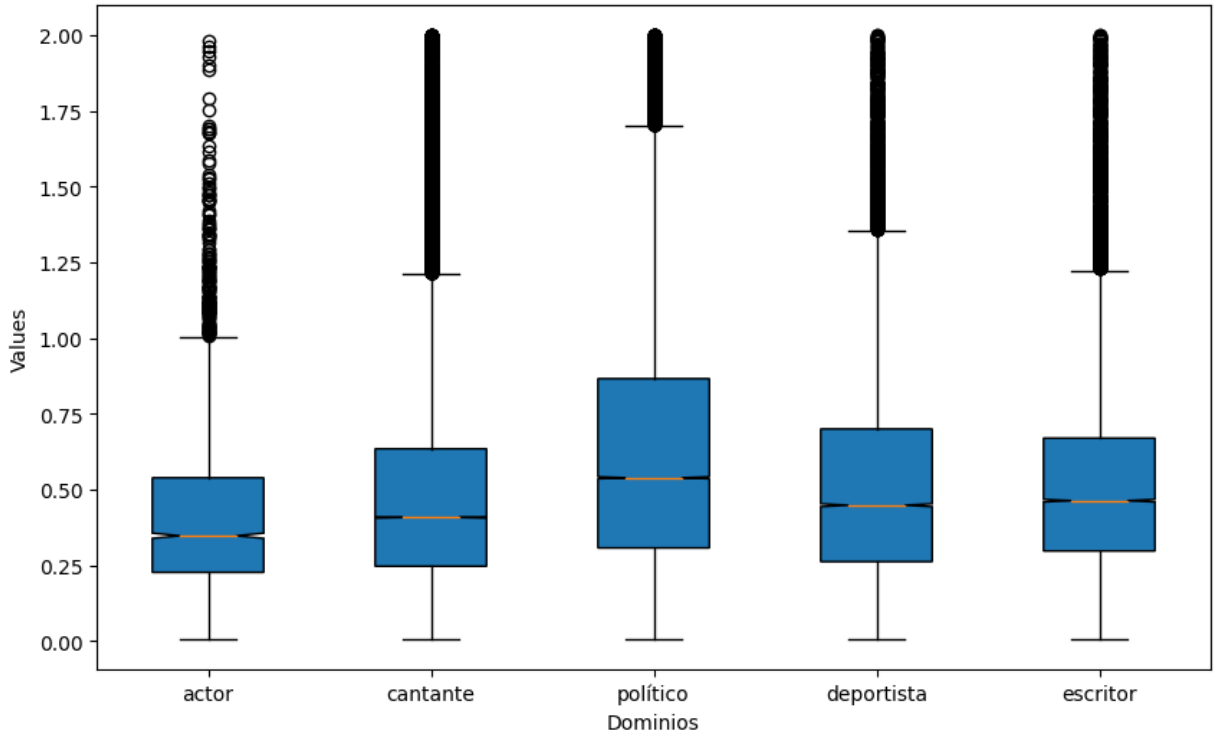


Figura 11. Distribución de la duración de pausas por dominio.

6. Discusión

6.1 Número De Pausas Encontradas Y Su Posible Error

En primer lugar, gracias al método utilizado se detectaron de manera automática 197.720 pausas en un corpus de 20.815 minutos de diálogos en entrevistas a celebridades, lo cual supera ampliamente el corpus usado en estudios anteriores tales como: el de Lundholm (2015), quien en su corpus PauDia analizó aproximadamente 60 minutos de grabaciones y quién también, utilizando una muestra de 563 segundos (~9 minutos), estudió el corpus SponTal (Edlund *et. al*, 2010). Otro ejemplo es el trabajo de Rochet-Capellan & Fuchs, S. (2014) donde se analizaron pausas en aproximadamente 275 minutos de diálogos.

Sin embargo, estimar la calidad de la identificación de estas pausas es una tarea difícil si se aborda comparando las pausas identificadas automáticamente con pausas identificadas manualmente por profesionales, debido a la magnitud del corpus utilizado, así como a los recursos destinados para esta investigación. Como alternativa es posible comparar la rata de pausas silenciosas encontradas en este estudio, que es de 9,5 pausas por minuto, contra la rata de producción de pausas encontrada por Lundholm en el corpus SponTal el cual fue de 18,2 pausas por minuto. En ese sentido, se puede estimar que el método utilizado permite identificar un 52,12% de las pausas, lo que se podría considerar como estimación del valor máximo de cubrimiento (*recall*) en el presente estudio.

Por su parte, Lundholm evaluó el método automático de detección de pausas usando el programa Praat (Boersma, P y Weenik, D, 2014), el cual obtuvo un cubrimiento máximo del 32% usando un umbral de detección óptimo de 45 dB ajustado manualmente. De este modo, es posible estimar que el cubrimiento del método presentado podría estar en un rango entre el 32% y el 52,12%. No obstante, hay que señalar que el método implementado en Praat se basa únicamente en la energía de la señal, ignorando el análisis espectral de las pausas, motivo por el cual se puede considerar que el cubrimiento de nuestro estudio está más cercano al valor superior del rango descrito. Esta idea cobra sentido cuando se observa que, en el campo del procesamiento de señales, se ha demostrado que la combinación de la energía de la señal y los centros de gravedad espectral son un método que mejora significativamente la detección de actividad de voz, en comparación con sólo el uso de la energía de la señal (Cen, L *et. al*, 2011).

De hecho, en el trabajo de Cen, L *et. al* (2011) el método de Giannakopoulos (2009) es usado como línea de base para comparar los resultados en la tarea de remoción de silencios en audios de habla, encontrando que para el corpus usado en esa investigación este método obtuvo una rata de error del 30,59%, lo que significa que la precisión es comparable en este estudio, ya que realizaron una tarea similar a la presentada en nuestro estudio y con datos parecidos a los que fueron utilizados en nuestro corpus de entrevistas. En ese sentido, resulta factible estimar la precisión del método de Giannakopoulos en el 69,41%. Sin embargo, al igual que en el caso del cubrimiento, se espera que esta precisión sea superior en el método presentado, dado que una buena cantidad de pausas detectadas con el método de identificación de señal de audio son descartadas gracias a la información proveída por Whisper. Además, los 4 casos seleccionados permiten descartar pausas donde no hay coincidencias claras entre los dos métodos combinados; sin mencionar que el método de Giannakopoulos ha sido adoptado en diferentes estudios recientes asumiendo una rata de error mucho menor que la reportada por Cen et al. (Sreekumar *et. al*.,2014; Godel *et. al*, 2023). De este modo, es razonable afirmar que una precisión de 69,41% es una estimación pesimista para el método implementado en este estudio.

En cualquier caso, asumiendo la estimación optimista del cubrimiento (52,12%) junto con la pesimista de la precisión (69,41%) se obtiene una estimación para la medida F1 de 59,54%, la cual al compararse con el mejor valor de esta medida observada por Lundholm (2015), el cual fue del 25%, representa una mejora considerable sobre el rendimiento de Praat en la tarea de identificación de pausas silenciosas, alcanzando aproximadamente una mejora del 138%. De igual manera, al comparar el valor de la estimación para la medida F1 del 59,54% con el acuerdo entre los dos anotadores del estudio de Lundholm, que fue de 0.674 usando la medida Kappa de Cohen, se muestra que el método utilizado en esta investigación se acerca al rendimiento de los humanos anotando manualmente las pausas. En ese sentido, los resultados alcanzados con el método descrito en el presente estudio permiten contradecir la conclusión de Lundholm (p.53) quien declara que las pausas identificadas automáticamente no son útiles para la investigación por su alta rata de error.

6.2 Análisis de Casos

En cuanto a la productividad de los 4 casos seleccionados (ver Figura 7) se observa que el Caso #2 resultó ser el que más pausas identificó, ya que es el que permite que haya una coincidencia del silencio detectado por el método de Giannakopoulos y las dos marcas de tiempo detectadas por Whisper-timestamped. Además, el Caso #2 supera ampliamente en productividad al Caso #1 donde se requiere un pequeño umbral entre los dos extremos izquierdo y derecho de las marcas de tiempo, de modo que el silencio detectado esté completamente incluido entre las marcas de tiempo. El segundo caso más productivo fue el Caso #3 donde el umbral solo se permite en el extremo derecho de la pausa, el cual identificó más de 6 veces la cantidad de pausas que su equivalente simétrico, el Caso #4. Esta asimetría puede ser explicada debido a la naturaleza de las anotaciones de Whisper-timestamped que buscan una coincidencia entre la pronunciación de las palabras y la presentación del subtítulo en el video, posiblemente porque las marcas de tiempo están diseñadas para que la palabra aparezca en el video justo cuando empieza su pronunciación, pero su marca de terminación estaría extendida en el tiempo algunos milisegundos para facilitar su lectura antes de desaparecer del video. Además, como se observa en las figuras **5.1 y 5.2,** el método de Giannakopoulos tiende a detectar el segmento de silencio de manera anticipada y termina la detección un poco prematuramente, contribuyendo a la asimetría en la efectividad de los casos #3 y #4.

6.3 Frecuencias Léxicas

Gracias al elevado número de pausas detectadas (cerca de 200.000), es posible obtener las frecuencias considerables de las palabras presentes tanto antes como después de dichas pausas, con el fin de realizar, por primera vez en el ámbito científico, un análisis de las frecuencias léxicas. Un fenómeno que se aprecia es que las frecuencias de las palabras antes de las pausas son considerablemente menores que las frecuencias de estas mismas después de las pausas (Ver tabla 2). Por ejemplo, la palabra “que”, la cual es la más frecuente del corpus, ocurre antes de las pausas 3.679 veces, mientras que aparece 12.286 veces después de la pausa. Este patrón se cumple en todos los casos posibles. Asimismo, se observa que la columna “Palabra antes” tiene solo 10 coincidencias con las 20 palabras más frecuentes del corpus, mientras que “Palabra después” tiene 17 coincidencias.

Teniendo en cuenta que la pausa puede ser utilizada como una estrategia para recuperar la siguiente palabra en el discurso, es posible que las palabras que cumplen la función gramatical de conjunción sean con mayor frecuencia utilizadas después de una pausa como estrategia adicional para mantener el turno de conversación y al mismo tiempo planificar la frase que viene a continuación. En cuanto a la alta frecuencia después de las pausas de palabras como "que", "de", "en", etc. es posible que sea producto de su función para conectar ideas, enfatizar información y llenar vacíos en el discurso, lo cual coincide con la función de la pausa para contrastar, presentar o focalizar la información.

Ahora bien, las ideas se pueden manifestar con mayor frecuencia si se inician con una palabra funcional, mientras que sería menos frecuente que se terminen con una palabra funcional. Es decir, al final de una frase hay una mayor posibilidad de usar una pausa para marcar el cambio de idea y preparar la siguiente frase puesto que palabras que cumplen funciones gramaticales como "que" o "y", inician la nueva frase después de la pausa, lo cual sería una posible explicación al fenómeno observado de una mayor frecuencia de palabras con función gramatical de conjunción después de la pausa.

El primer patrón encontrado como más frecuente en los bigramas antes y después de la pausa son las repeticiones “no [Pausa] no”, “sí [Pausa] sí”, junto a “que [Pausa] es”, lo cual puede deberse al uso de muletillas por parte de los hablantes, ya que la naturaleza de las pausas es que son interrupciones que surgen de manera espontánea en el discurso y que en la mayoría de los casos no son planeadas (Machuca *et al.*, 2015; Williams, 2023). Esto significa que la repetición de palabras podría ser una consecuencia de la improvisación de los hablantes.

Otro de los casos frecuentes como “es [Pausa] que”, “lo [Pausa] que”, “cosas [Pausa] que” y “gente [Pausa] que”, la palabra “que” tiene una función catafórica. Es posible que la presencia de la pausa sea por el tiempo que requiere el hablante para recuperar la expresión que resuelve la catáfora más adelante en el discurso. En los casos “años [Pausa] y”, “bien [Pausa] y”, “vida [Pausa] y”, “eso [Pausa] y”, “país [Pausa] y” se observa la aparición sistemática de la palabra “y” luego de la pausa, lo que se puede explicar por la función gramatical de conjunción para presentar ideas, lo que a su vez coincide con la mayor frecuencia reportada de “y” con 19.068 en la palabra después de las pausas.

6.4 Distribución de la Duración de las Pausas

Al observar la distribución de las pausas silenciosas encontradas en nuestro corpus de celebridades, se observa mayor frecuencia en las pausas que se encuentran alrededor de los 250 ms (ver Figura 8). Cuando son mayores de este valor, disminuye sistemáticamente la frecuencia (conteo) a medida que aumenta la duración hasta llegar a una frecuencia muy baja en la duración de 2 segundos. Por otro lado, las pausas inferiores a los 250 ms decrecen rápidamente en frecuencia acercándose al límite de pausa mínima de 150 ms, lo que coincide con las apreciaciones sobre las pausas mínimas percibibles de Lundholm (p. 38). Al comparar nuestro resultado con trabajos realizados para el inglés y el francés realizado por Grosjean y Deschamps (1975, p. 158) se observa gran similitud en la distribución a pesar de la enorme diferencia de datos utilizados, lo cual confirma la validez de las pausas identificadas automáticamente en nuestro estudio.

6.5 Distribución Del Número De Sílabas En Las Palabras Alrededor De Las Pausas

En la figura 9 se evidencia un patrón sistemático en el cual las palabras antes de las pausas silenciosas tienen más sílabas que las palabras después. Este fenómeno podría estar relacionado con el hecho de que, como se explicó anteriormente, las palabras después de las pausas son de alta frecuencia, y a su vez, se relacionan con palabras de pocas sílabas. Asimismo, en línea con lo señalado por Zellner (1994) y Lundholm (2015) las pausas pueden ser motivadas por la necesidad fisiológica de tomar aire para continuar con el discurso, por lo que seguramente es necesario hacer una pausa luego de producir palabras con más sílabas, mientras que luego de la pausa se utilizan palabras con menos sílabas para continuar con la conversación. No obstante, las pausas también son producto de una estrategia cognitiva para recuperar una idea sin perder el turno de conversación, y al considerar que las palabras más largas suelen ser más complejas, es decir, demandan mayor procesamiento cognitivo, es probable que luego de ser pronunciadas los hablantes necesiten de un respiro que facilite tanto el procesamiento de la información como la comprensión del oyente.

Cuando se revisan trabajos relacionados con las pausas silenciosas, el patrón de que las palabras antes de las pausas tengan sistemáticamente más sílabas resulta ser una observación nueva para el español, lo cual amerita estudios posteriores relacionados con la prosodia.

6.6 Distribución De La Duración De Las Pausas Según Función Intermedia O Gramatical

Como se puede apreciar en la figura 10, las pausas gramaticales son considerablemente más largas que las intermedias, lo que se podría explicar debido a que cuando el hablante tiene la necesidad de hacer una pausa en medio de una frontera gramatical, se produce un efecto de prolongación de la pausa. Siguiendo a Ferreira (1991), este fenómeno es de orden cognitivo ya que, al terminar una idea, los hablantes comienzan a componer otra, lo cual llevaría más tiempo si el sujeto y el objeto son sintácticamente más complejos. Ahora bien, en sentido estricto, Ferreira considera la complejidad en las estructuras internas de las oraciones, mientras que en este estudio abordamos la complejidad nivel de frases, no obstante, en ambos casos se confirma la relación entre una mayor duración de las pausas silenciosas cuando hay una tarea cognitiva de mayor dificultad. En la perspectiva de Horne *et al.* (1995) el alargamiento de las pausas se debe a la presencia de un fuerte límite prosódico. Por el contrario, las pausas intermedias podrían ser más cortas debido a que no buscan contrastar o enfatizar información nueva, sino que se producirían principalmente por una necesidad fisiológica de tomar aire, pasar saliva, o generar expectativa a su interlocutor.

Se observa que la distribución de las duraciones de las pausas gramaticales encontradas en nuestro estudio tiene una alta coincidencia con la distribución reportada por Lundholm (2015, p. 44, Figura 3.1.), a pesar de que en su investigación no se diferenciaron las pausas gramaticales de las intermedias.

6.7 Acentuación Léxica Alrededor De Las Pausas Silenciosas

En la tabla 3 se aprecia un patrón de aparición predominante de palabras graves antes de las pausas siendo el 61.63% de los casos, el cual llama la atención si se compara con la proporción de las palabras graves en todo el corpus de celebridades, que es de 40.53%. En otras palabras, se observa una tendencia clara por encima del patrón aleatorio. Hay que señalar que este fenómeno se encuentra más presente en las pausas gramaticales (antes de una coma) que en las pausas intermedias, aunque en ambos casos es bastante frecuente. Es importante señalar que, al hacer una revisión corta de la literatura relevante en español, no encontramos evidencia alguna de observaciones previas de este fenómeno.

De acuerdo con lo observado en la sección 6.5, las palabras antes de la pausa tienen más sílabas que las de después, lo que indica que son palabras más largas. Ahora bien, se evidencia además que son palabras

predominantemente graves, lo que podría sugerir que su aparición sistemática hace que los hablantes las perciban como más prominentes dentro del flujo de habla. Es posible que las pausas intermedias, siendo precedidas de una prominencia auditiva (tanto por la duración como por el acento grave de las palabras antes de la pausa) cumplan con la función de llamar la atención o resaltar un momento particular en el discurso.

En cuanto a su acentuación y duración combinada, las palabras graves y largas, al tener mayor variedad de fonemas, junto a una mayor complejidad desde el punto de vista articulatorio que las palabras agudas o esdrújulas, podrían presentar una mayor prominencia auditiva. Esto sugeriría que en las pausas el patrón más frecuente de transición "grave (muchas sílabas) [pausa] aguda (pocas sílabas)" estaría justificado por la prominencia auditiva de la palabra antes, mientras que la palabra después estaría cumpliendo con el patrón aleatorio de que las palabras en el discurso en español son en su mayoría agudas,como se evidencia en el caso palabras monosílabas que son usualmente funcionales en la tabla 2.

En la tabla se observan otros fenómenos en casos como las variaciones de los conteos según si la pausa es intermedia o gramatical. Sin embargo, el análisis detallado de dichos conteos excede los alcances de este estudio y se perfila como perspectivas de investigación futuras.

6.8 Efecto De Frecuencia De Las Palabras Y Duración De Las Pausas

La correlación entre la frecuencia de palabra antes de pausa en el corpus frente a la duración de la pausa es de $\rho = -0.165$, lo que revela que cuando el proceso cognitivo de recuperación de la palabra previa a la pausa es eficiente, por word-frequency effect, entonces se relaciona con economía en la duración de la pausa. En ese sentido, se podría afirmar que el proceso cognitivo de producción del habla presenta una continuidad cognitiva cuando hay dos procesos cognitivos eficientes: primero, la recuperación rápida de la palabra antes de la pausa debido a su alta frecuencia, y segundo, la producción de una pausa corta.

Esta hipótesis se reforzaría con la correlación existente entre el número de letras en la palabra antes de pausa y duración de la pausa de $\rho = 0.172$. La razón, es que una alta cantidad de letras de una palabra se podría asociar con una mayor complejidad silábica, es decir, con una carga cognitiva alta, lo que estaría relacionada con pausas silenciosas largas. Se observa además, que las otras dos correlaciones reportadas en la tabla 4

asociadas a la palabra después son de menor efecto, lo que sugiere que son los factores cognitivos de la palabra antes, aquellos que determinan en mayor medida la duración de la pausa.

Sobre la correlación entre la frecuencia en el corpus de la transición entre la palabra antes y después frente a la duración de la pausa ($\rho = -0.097$), se evidencia que el proceso cognitivo de la recuperación de las palabras antes y después de la pausa tiene una relación con la duración de la pausa. Esto significa que procesos eficientes de recuperación asociados a frecuencias altas en el corpus guardan relación con pausas cortas, es decir, económicas o eficientes. Teniendo en cuenta esta idea, se refuerza la hipótesis de que los procesos cognitivos llevan un ritmo uniforme en el discurso.

Asimismo, la correlación entre la velocidad media de producción de las palabras y la duración de las pausas es de $\rho = -0.416$, lo cual refuerza la idea de que un ritmo rápido en el habla implica procesos cognitivos eficientes, es decir, con pausas de corta duración. Este resultado confirma el trabajo de Kendall (2009) quien desde una perspectiva sociolingüística estudió la velocidad de articulación de los habitantes, encontrando así una tendencia de las pausas a ser más cortas cuando esta velocidad era menor.

6.9 Análisis Por Dominios

De acuerdo con lo presentado en la figura 11, los dominios de las celebridades se pueden ordenar según la mediana de la duración de las pausas silenciosas de la siguiente manera: 1) políticos, 2) escritores, 3) deportistas, 4) cantantes y 5) actores; siendo uno el de mayor duración y cinco el de menor duración.

Este orden podría estar asociado con la posible complejidad de los temas en los diálogos. En ese sentido, los políticos, representados en nuestro corpus por tres jefes de estado, abordan en sus diálogos temas de gran importancia social, ya que su rol como presidentes implica un mayor cuidado y preparación en sus entrevistas. En segundo lugar, se encuentra el dominio de los escritores que representan una comunidad erudita cuyas manifestaciones tienen un grado importante de complejidad. En tercer lugar, se encuentran los deportistas, quienes tratan temas de carácter técnico relacionado con su deporte, así como con los torneos competitivos donde participan, es decir, presentan una complejidad media en sus diálogos. Finalmente, los

actores y cantantes, al formar parte del dominio del entretenimiento, generalmente tienen diálogos informales sobre temas personales o profesionales, lo cual lo ubicaría en un nivel de baja complejidad.

Es posible afirmar entonces que existe una relación clara entre la complejidad del discurso y la longitud de las pausas. En estudios relacionados con el aprendizaje de una segunda lengua, se evalúa la precisión, fluidez y complejidad como criterios para determinar el dominio de las personas sobre dicha lengua. Se ha observado que las pausas silenciosas son más frecuentes cuando el hablante tiene una competencia lingüística menor, puesto que necesita el tiempo adicional para organizar sus ideas en la segunda lengua (Mora y Ferrer, 2012), así como se puede apreciar que las pausas silenciosas aparecen con mayor frecuencia cuando el contexto de comunicación tiene una complejidad cognitiva mayor. (Williams, 2023).

En investigaciones como la de Rodríguez (2013), la de Potagas et.al (2022), la de Balogh et.al (2023), o la de Angelopoulou *et al.* (2024) se estudia la relación entre la duración de las pausas y el estado cognitivo de los hablantes, puesto que en ambos trabajos la población fueron personas con afasia a quienes se les solicitaba llevar a cabo distintas tareas lingüísticas con el fin de evaluar la duración y producción de pausas cuando se requerían tareas más complejas, llegando a concluir que la presencia de las pausas silenciosas pueden contribuir a la identificación de deterioros cognitivos leves en la fluidez de los hablantes en etapas tempranas de afasia . Los trabajos concluyen que no hay diferencias entre la producción de pausas entre personas sanas y personas con afasia, aunque en estas últimas sí se evidencia una duración mayor.

En resumen, la relación entre la complejidad cognitiva y la presencia de pausas silenciosas ha sido estudiada en el aprendizaje de segunda lengua, así como en el análisis de pacientes con afasia quienes presentan dificultades cognitivas para cumplir con tareas de mayor dificultad. Sin embargo, según lo observado en este estudio en la duración de las pausas por dominio, resulta evidente que existe una complejidad cognitiva asociada a la "sofisticación" de los temas para los hablantes sin condiciones médicas de afasia, haciendo uso de su propia lengua; en otras palabras, los temas con mayor implicación social están relacionados con pausas más largas, debido a que hay una alta dificultad cognitiva en aquellos diálogos donde se sanciona con mayor severidad la postura del hablante, quien toma un tiempo mayor para planear su discurso.

6.10 Limitaciones

En primer lugar, la herramienta Whisper-timestamped no podía hacer *speaker-diarization* (identificación de hablante), motivo por el cual no fue posible distinguir los hablantes en los diálogos. En ese sentido, para garantizar que las pausas estudiadas no estuvieran asociadas a cambios de turno de los hablantes, sólo se consideraron las pausas intermedias y las marcadas con una coma. Sin embargo, el método puede verse perjudicado en términos de cubrimiento, ya que posiblemente algunas pausas marcadas con "." o "?" no necesariamente implican un cambio de turno y fueron ignoradas en este estudio. La limitación se puede superar usando herramientas como WhisperX (Bain et al., 2023) que logran la identificación de los hablantes y los cambios de turno, aunque requieren recursos computacionales de alto costo.

Por otra parte, en el presente estudio se deja de lado el análisis de estructuras previas y posteriores más largas tales como enunciados, oraciones o frases debido a las limitaciones de tiempo y a la magnitud del corpus escogido, a saber, entrevistas a celebridades del mundo hispanohablante.

Asimismo, la determinación de la precisión, cubrimiento, y medida F1 son estimaciones fundamentadas en razonamiento, junto a la comparación con resultados de trabajos similares. En futuras investigaciones que usen esta metodología, se deberían calcular estas medidas a través de una cantidad considerable de anotaciones manuales de las pausas. De este modo, se podría hacer una comparación frente a la gran cantidad de pausas encontradas con el método propuesto en este estudio.

7. Conclusiones

Los resultados obtenidos a partir de la aplicación de la nueva metodología para detección automática de pausas silenciosas sugieren que la complejidad del discurso, tanto a nivel cognitivo como en diferentes niveles de la lengua, se relaciona con la duración de las pausas, lo que sugiere que entre mayor sea la complejidad cognitiva de la producción y planeación del discurso hablado, mayor será la duración de la pausa.

Esta afirmación está soportada por varios resultados de este estudio. El primero, es que a nivel morfológico-fonológico se observó el fenómeno de predominancia de mayor cantidad de sílabas y letras en la palabra antes de la pausa, lo cual indica una mayor complejidad fonológica en el antecedente de la pausa que en la palabra subsiguiente. Además se evidenció una relación entre el número de letras de la palabra que precede la pausa frente con la duración de la pausa, puesto que cuando la palabra tiene más letras, se observa que la duración de la pausa es mayor posiblemente asociado a la complejidad articulatoria.

En segundo lugar, se observó que las pausas silenciosas gramaticales tienen una duración mayor que las intermedias, lo que se relaciona con el proceso cognitivo de planeación gramatical el cual agrega complejidad en el discurso. En tercer lugar, a nivel entonativo (o prosódico-fonológico) las palabras graves resultaron ser predominantes cuando están precediendo a las pausas, lo cual llama la atención si se considera que estas palabras son de mayor complejidad articulatoria por la doble variación del tono en las palabras de más de tres sílabas. En cuanto al cuarto factor, cuando se analizan las frecuencias de aparición en el corpus de las palabras asociadas con la eficiencia y complejidad tanto de recuperación como planeación, se puede apreciar que cuando más baja es la eficiencia y mayor es la complejidad de las palabras (bajas frecuencias en el corpus), es más común que las pausas sean más largas. Finalmente, la complejidad de la tarea del discurso, asociada con la complejidad de los temas de los diferentes dominios estudiados, evidencia también una relación directa, lo que significa que temas más complejos (política) se asocian a pausas con mayor duración en comparación con temas menos complejos (farándula).

En conclusión, estas cinco evidencias muestran que la complejidad del discurso, a nivel morfológico-fonético (sílabas y palabras), gramatical (comas), prosódico (sílabas tónicas), cognitivo (frecuencias de eficiencia y recuperación), así como semántico (complejidad en los temas), tiene un vínculo directo con la duración de las pausas silenciosas. Esta relación entre complejidad y duración de las pausas es soportada con evidencia múltiple por primera vez en este estudio.

Vale la pena señalar que el análisis de frecuencias léxicas en las palabras alrededor de las pausas silenciosas, permite evidenciar patrones interesantes que ameritan investigación futura. Sin embargo, dichos patrones no aportan evidencia clara que contribuya a soportar el hallazgo principal de este estudio.

Por último, hay que decir que la metodología utilizada en este estudio permite, por una parte, validar como recurso de investigación en lingüística a los videos de YouTube, los cuales suelen ser descartados por la heterogénea calidad de los audios; y por otra, da crédito al uso de métodos de detección automática ya que permite encontrar tendencias claras, con baja varianza, considerando que los resultados son similares y comparables con otras investigaciones. Concluimos, que el uso de herramientas automáticas para la anotación de pausas silenciosas permite analizar corpus de gran tamaño y observar fenómenos no percibibles en corpus pequeños.

Referencias bibliográficas

- Angelopoulou, G.,Kasselimis, D., Varkanitsa, M., Tsolakopoulos, D., Papageorgiou, G., Velonakis, G., Meier, E., Karavassilis, E., Pantoleon, V., Laskaris, N., Kelekis, N., Tountopoulou, A., Vassilopoulou, S., Goutsos, D., Kiran, S., Weiller, C., Rijntjes, M., Potagas, C. (2024). *Investigating silent pauses in connected speech: integrating linguistic, neuropsychological, and neuroanatomical perspectives across narrative tasks in post-stroke aphasia.* Frontiers in Neurology (15). 10.3389/fneur.2024.1347514.

- Bain, M., Huh, J., Han, T., Zisserman, A. (2023) WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Conferencia INTERSPEECH. Dublin.

- Balogh, R., Imre, N., Gosztolya, G., Pákáski, M., & Kálmán, J. (2023). *The role of silence in verbal fluency tasks–a new approach for the detection of mild cognitive impairment*. Journal of the International Neuropsychological Society, 29(1), 46-58.

- Blondet, M. (1999). *Estudio acústico-prosódico de los fenómenos sonoros de hesitación: análisis contrastivo entre los dialectos andino y central*. (Tesis de Maestría. Universidad de los Andes, Mérida, Venezuela).

- Blondet, M., (2006). *Variaciones de la velocidad de habla en español: patrones fonéticos y estrategias fonológicas. Un estudio desde la producción*. (Tesis doctoral. Universidad de los Andes,Mérida, Venezuela).

- Boersma, P. y Weenink, D. (2014). Praat: doing phonetics by computer. Computer program.

- Borzi, C., Trípodi, M., y García Jurado, M. A. (2017). *Confluencia entre pistas perceptivas y configuraciones prosódicas del marcador discursivo "entonces" en posición intercláusulas*. Exploraciones fonolingüísticas. *V Jornadas Internacionales de Fonética y Fonología y I Jornadas Nacionales de Fonética y Discurso*, 219-232.

.

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

.

- Calsamiglia, H. y H.Tusón (2007). *Las cosas del decir.* Manual de análisis del discurso, Barcelona, Ariel.

Clark, H. H. (2006). *Pauses and hesitations: Psycholinguistic approach.* In K. Brown (Ed.), Encyclopedia of language and linguistics, (1), 284,288.

Cucchiarini, C., Strik, H., & Boves, L. (2002). *Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech*. The Journal of the Acoustical Society of America, 111(6), 2862–2873.

Curhan, J. R., Overbeck, J. R., Cho, Y., Zhang, T., & Yang, Y. (2022). *Silence is golden: Extended silence, deliberative mindset, and value creation in negotiation*. Journal of applied psychology, 107(1), 78.

- Cen, L., Dong, M., & Chan, P. (2011). *Segmentation of speech signals in template-based speech to singing conversion*. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (pp. 1-4).

- Crible, L., Degand, L., Gilquin, G. (2017). *The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis) fluency*. Languages in Contrast, 17(1), 69-95.

- Dall, R., Tomalin, M., Wester, M., Byrne, W., y King, S. (2015). *Investigating Automatic y Human Filled Pause Insertion for Speech Synthesis*. Proceedings of Interspeech, 2015, 3002-3006. https://www.isca-speech.org/archive/interspeech_2015/i15_3002.html

- Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., y House, D. (2010). Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture. In LREC, 2992–2995

- Ferreira, F. (1991). *Effects of length and syntactic complexity on initiation times for prepared utterances*. Journal of Memory and Language, 30(2), 210-233.

- Giannakopoulos, T. (2009). A method for silence removal and segmentation of speech signals, implemented in Matlab. University of Athens, Athens, 2, 17.

.

- Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. Journal of statistical Software, 31, 1-24.

- Godel, M., Robain, F., Journal, F., Kojovic, N., Latrèche, K., Dehaene-Lambertz, G., y Schaer, M. (2023). Prosodic signatures of ASD severity and developmental delay in preschoolers. NPJ Digital Medicine, 6(1), 99.

- Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction.* Language and Linguistics Compass, 3(1), 397-417. https://doi.org/10.1111/j.1749-818x.2008.00128.x

- Grosjean, F., Deschamps, A. (1975). *Analyse contrastive des variables temporelles de lánglais et du francais: vitesse de parole t variables composantes, phenomenes d´hestiation.* Phonetica 31, 144-18.

- Horne, M., Strangert, E., y Heldner, M. (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. Proceedings ICPhS, (3).

Housen, A., & Kuiken, F. (2009). *Complexity, accuracy, and fluency in second language acquisition*. Applied linguistics, 30(4), 461-473.

- Kendall, T. (2009). Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project. PhD thesis, Duke University.

- Labov, W. (1972). S*ome Principles of Linguistic Methodology in Language.Society*, (1), 97–120. Recuperado de: http://www.jstor.org/stable/4166672

- Louradour, J. (2023). whisper-timestamped [Computer software]. GitHub repository. https://github.com/linto-ai/whisper-timestamped

- Lundholm, K (2015). Production and perception of pauses in speech. phD Tesis, Gothenburg University.

- Machuca (2018). *Pausas sonoras y bilingüismo.* Revista de Estudios de Fonética Experimental, pp.75-95.

- Machuca, M. J.; J.Llisterri y A. Ríos (2015). *Las pausas sonoras y los alargamientos en español: un estudio preliminar*, Normas. Revista de Estudios Lingüísticos Hispánicos, 5, 81-96.

- Martín, A., González-Carrasco, I., Rodriguez-Fernandez, V., Souto-Rico, M., Camacho, D., y Ruiz-Mezcua, B. (2021). *Deep-Sync: A novel deep learning-based tool for semantic-aware subtitling synchronisation*. Neural Computing and Applications, 1-15.

- Mirzaei, M. S., Meshgi, K., Akita, Y., y Kawahara, T. (2017). Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. ReCALL, 29(2), 178-199.

Mora, J. C., & Valls-Ferrer, M. (2012). *Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts*. Tesol Quarterly, 46(4), 610-641.

- Pálvölgyi, K. (2020). *The duration of filled pauses and prolongations in northern and southern dialects of Spanish*. Revista de Estudos Linguísticos da Univerdade do Porto, 15, 71–93.

Potagas, C., Nikitopoulou, Z., Angelopoulou, G., Kasselimis, D., Laskaris, N., Kourtidou, E., ... & Kapaki, E. (2022). *Silent pauses and speech indices as biomarkers for primary progressive aphasia*. Medicina, 58(10), 1352.

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

- Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130399.

- Rodríguez, J. (2013). *Las pausas en el discurso de individuos con demencia tipo Alzheimer*. Estudio de casos. Lengua y Habla, (17), 253-267.

- Rose, R. (2017). Silent and filled pauses and speech planning in first and second language production. Proceedings of DiSS, 2017, 49-52·

- Sacks, H., Schegloff, E., & Jefferson, G. (1974). *A simplest systematics for the organization of turn-taking for conversation*. Language, 50(4):696–735.

Schegloff, E. A. (1992). *To Searle on conversation: A note in return*. H. Parret & J. Verschueren (Eds.), 113-128.

Schegloff, E. A. (2003). *Conversation analysis and communication disorders*. Conversation and brain damage, 21-55.

- Sreekumar, K. T., George, K. K., Arunraj, K., & Kumar, C. S. (2014, January). Spectral matching based voice activity detector for improved speaker recognition. In 2014 International Conference on Power Signals Control and Computations (EPSCICON) (pp. 1-4). IEEE.

- Stenström, A. (1990). *Pauses in Monologue and Dialogue*. The London-Lund Corpus of Spoken English: Description and Research, J. Svartvik (ed.). Lund: Lund University Press, 211–252.

Tannen, D. (1995). *The power of talk: Who gets heard and why*. Harvard Business Review, 138-148.

- Tottie, G. (2017). *From pause to word: uh, um and er in written American English*. English Language and Linguistics, 23(1), 105-130.

- Wang Y., Jordan R., Ignatius S.B. Nip; Ray D. & Kent, J.(2010). *Breath Group Analysis for Reading and Spontaneous Speech in Healthy Adults.* Folia Phoniatr Logo 62 (6): 297–302. https://doi.org/10.1159/000316976.

- Williams, S. (2023). *Silents pauses. Repetitions*. Disfluency and proficiency in second language speech production. 31-72, 147-174. https://doi.org/10.1007/978-3-031-12488-4

- Zellner, B. (1994). *Pauses and the temporal structure of speech*. E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley: .41-62.