Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 209-220 recibido 03-04-2024 revisado 09-05-2024 aceptado 17-05-2024

Adaptación de ASR al habla de personas con síndrome de Down

ASR model adaptation to the speech of people with Down syndrome

David Fernández-García,[1] **Valentín Cardeñoso-Payo,**[1]
César González-Ferreras,[1] **David Escudero-Mancebo**[1]
[1]Grupo de Investigación ECA-SIMM, Universidad de Valladolid, España
david.fernandez@estudiantes.uva.es,
{valentin.cardenoso, cesargf, escuderomancebo.david}@uva.es

Resumen: El habla de las personas con discapacidad intelectual (DI) plantea enormes retos a los sistemas de reconocimiento automático del habla (ASR), dificultando con ello el acceso de una población especialmente sensible a los servicios de información. En este trabajo se estudian las dificultades de los sistemas ASR para reconocer habla de personas DI y se muestra cómo esta limitación puede ser combatida con estrategias de ajuste fino de modelos. Se mide el rendimiento de ASR basado en *whisper* (v2 y v3) con un corpus de referencia de habla típica y habla DI, comprobando que hay diferencias importantes y significativas. Aplicando técnicas de *fine-tuning*, el rendimiento para hablantes DI mejora en al menos 30 puntos porcentuales. Nuestros resultados muestran que la inclusión de voz de personas DI en los corpus de entrenamiento es fundamental para mejorar la eficacia de los ASR.
Palabras clave: ASR, Habla anómala, *whisper*, Aumento de datos.

Abstract: The speech of people with intellectual disabilities (ID) poses enormous challenges to automatic speech recognition (ASR) systems, making it difficult for a particularly sensitive population to access information services. This work studies the difficulties of ASR systems in recognizing the speech of ID people and shows how this limitation can be combated with model fine-tuning strategies. The performance of ASR based on *whisper* (v2 and v3) is measured with a reference corpus of typical speech and DI speech, verifying that there are important and significant differences. By applying *fine-tuning* techniques, performance for DI speakers improves by at least 30 percentage points. Our results show that the inclusion of the voice of ID people in the training corpora is essential to improve the effectiveness of ASRs.
Keywords: ASR, Pathologic Speech, *whisper*, Data Augmentation.

1 Introducción

La calidad de los sistemas de reconocimiento automático de habla ha aumentado significativamente en los últimos años, lo que ha mejorado sustancialmente la accesibilidad de las aplicaciones por medio de interfaces habladas. Sin embargo, es ya conocido que su eficacia y precisión pueden variar significativamente entre diferentes grupos demográficos y condiciones de habla (Lea et al., 2023; Cibrian et al., 2024; De Russis y Corno, 2019). En concreto, su aplicación a poblaciones específicas como las personas con síndrome de Down, presenta una serie de dificultades, asociadas tanto a las limitaciones físicas como cognitivas de este tipo de personas (Hu et al., 2013).

El síndrome de Down (SD), según el nuevo sistema de clasificación de diagnóstico DSM-5 (American Psychiatric Association, 2013), es un subtipo de trastorno del desarrollo intelectual caracterizado por importantes limitaciones tanto en el funcionamiento intelectual (con un coeficiente intelectual igual o inferior a 70, con dificultades cognitivas) como en la conducta adaptativa. Las personas con SD muestran dificultades en las habilidades adaptativas conceptuales, sociales y prácticas. Aunque existe una gran heterogeneidad dentro de este colectivo, generalmente presentan también importantes dificultades en sus habilidades lingüísticas.

ISSN 1135-5948 DOI 10.26342/2024-73-15

Aunque todas las áreas del lenguaje están afectadas en distinto grado, las habilidades de producción de lenguaje suelen estar más deterioradas que las habilidades de comprensión (Martin et al., 2009). Las personas con SD pueden mostrar problemas de articulación, y en algunos casos, su habla es casi ininteligible. La inteligibilidad del habla se ve seriamente afectada por la presencia de errores en la producción de algunos fonemas, la pérdida de consonantes y la simplificación de sílabas (Laws y Bishop, 2004; Wong et al., 2015).

Las personas con SD a menudo afrontan importantes limitaciones en sus relaciones sociales debido a un manejo deficiente de la comunicación oral (Cleland et al., 2010; Martin et al., 2009; Chapman, 1997). Es fundamental comprender estas dificultades para proporcionar un apoyo adecuado y fomentar la inclusión en la sociedad, especialmente considerando que las tecnologías de la información y la comunicación (TIC) son parte integral de nuestras actividades diarias, incluyendo a las personas con discapacidad intelectual (Tanis et al., 2012; Feng et al., 2010). Por ejemplo, las redes sociales, una de las herramientas de TIC más utilizadas, también son frecuentemente utilizadas por personas con discapacidad intelectual (Caton y Chapman, 2016).

Aunque estudios relativamente recientes muestran que menos del 4 % de los usuarios con SD usan sistemas de entrada vocal como modo de interacción con los dispositivos digitales (Feng et al., 2010), el reconocimiento automático del habla tiene el potencial de hacer que la tecnología sea más accesible para los usuarios, especialmente para los que, como éstos, pueden tener problemas motores que merman la destreza a la hora de manejar el teclado o el ratón (Hu et al., 2013).

Sin embargo, los sistemas de reconocimiento del habla actuales no proporcionan resultados de la misma calidad para personas con SD, en comparación con aquellas con desarrollo típico (Cibrian et al., 2024). Por ello, y en línea con trabajos anteriores (Shor et al., 2019; Green et al., 2021), la solución adoptada en este trabajo es la adaptación de los sistemas de reconocimiento de habla independientes de locutor de última generación para garantizar una mayor precisión del reconocedor y, con ella, una mejora de la accesibilidad.

En este trabajo analizaremos las estrategias de adaptación, inspirados en el trabajo de Tobin y Tomanek (2022), que maneja también datasets de tamaño reducido para realizar *fine-tuning* de modelos de reconocimiento, y cómo aquellas pueden mejorar la tasa de reconocimiento para este grupo de usuarios con SD.

El artículo comienza revisando el estado del arte sobre el reconocimiento de habla de personas con SD. En la sección tres se describen los corpus utilizados para el entrenamiento y la evaluación del modelo, tanto de habla Down como de habla típica. En la sección cuatro describimos la metodología empleada, incluyendo las técnicas de aumento de datos y de adaptación de los modelos *whisper* (Radford et al., 2023), que se han tomado como referencia al ser los de mejor rendimiento conocido en habla típica en el momento de escribir este artículo (Cibrian et al., 2024). En la sección cinco se analizan los resultados obtenidos, comparando el rendimiento del modelo en ambos tipos de habla, y se evalúa el impacto de las técnicas de adaptación y aumento de datos. El artículo finaliza con un resumen de las principales conclusiones y se proponen líneas de trabajo futuro para mejorar el ASR en personas con SD.

2 Estado del arte

El reconocimiento automático del habla para el habla patológica es un área de investigación en crecimiento. Varios estudios han resaltado la importancia del ASR para ayudar a individuos con trastornos del habla (Rosen y Yampolsky, 2000; Kitzing, Maier, y Åhlander, 2009; Schultz et al., 2021). Uno de los trastornos del habla más estudiados es la disartria (Almadhor et al., 2023; Janbakhshi, Kodrasi, y Bourlard, 2021; Jiao et al., 2018; Shahamiri, 2021; Bhat y Strik, 2020). Los estudios han explorado el uso de sistemas ASR para reconocer patrones de habla disártrica, con un enfoque centrado en el análisis de los factores que afectan el rendimiento del sistema. El estado actual del ASR para el habla patológica, en particular la disartria, está avanzando rápidamente, ofreciendo soluciones innovadoras para ayudar a individuos con trastornos del habla en sus procesos de evaluación y terapia. La integración de la tecnología ASR tiene un gran potencial para mejorar las vidas de aquellos afectados por condiciones de habla patológica.

El habla de los individuos con SD no es necesariamente disártrica (Kumin, 2012), pe-

ro en general no es un habla típica, debido a problemas relacionados con el tono muscular bajo, el tamaño grande de la lengua y las frecuentes infecciones del oído. Estos individuos pueden experimentar habitualmente retrasos en el habla y dificultades en la articulación cabe destacar que cada persona con SD es única, y los problemas de habla pueden variar ampliamente entre individuos (Kumin, 2012).

La disponibilidad de corpus de entrenamiento en ocasiones es un obstáculo para la investigación y el desarrollo de estos sistemas, porque recopilar este tipo de corpus es un proceso costoso al que hay que dedicar gran cantidad de recursos. En el marco del proyecto Euphonia,[1]desarrollado por Google Research, se ha recopilado un corpus de habla de personas con trastornos del habla en inglés. Se trata del corpus más grande de este tipo de hablantes grabado hasta la fecha, puesto que han participado más de 1000 personas y se han grabado más de 1 millón de locuciones, lo que supone más de 1300 horas. El corpus incluye grabaciones de 105 personas con SD, lo que supone el 18,1 % del corpus (MacDonald et al., 2021). Empleando este corpus se han realizado diversos experimentos usando diferentes técnicas de deep learning para clasificar la inteligibilidad del habla de 661 locutores con diversas patologías, incluyendo SD (Venugopalan et al., 2021). Por otro lado, para mejorar el rendimiento de los sistemas de reconocimiento automático del habla en hablantes con trastornos del habla, se ha propuesto la personalización de modelos (Tomanek et al., 2021; Green et al., 2021), como vía para evitar la degradación significativa del rendimiento en habla patológica que se produce en los actuales sistemas de reconocimiento automático de habla.

Otro problema que suele afectar a este tipo de habla son las disfluencias y las variaciones en la pronunciación del habla, que pueden degradar gravemente el rendimiento del reconocimiento de habla, ya que los sistemas actuales de ASR se entrenan principalmente con habla fluida de hablantes típicos. Un enfoque sencillo para mejorar el rendimiento es ajustar los parámetros de decodificación en un sistema de reconocimiento de habla existente, lo que puede mejorar la tasa de errores de palabras (WER) para personas con trastornos de fluidez (Mitra et al., 2021).

[1]https://sites.research.google/euphonia/about/

3 Corpora

Para el desarrollo del trabajo se han usado tres corpora: FLEURS (Conneau et al., 2023) como corpus de referencia de habla típica para ajustar los hiperparámetros de *fine-tuning*, PRAUTOCAL Down (Escudero-Mancebo et al., 2022) como corpus de referencia de habla Down, para uso en *fine-tuning* y evaluación, y PRAUTOCAL Típico y VoxPopuli (Wang et al., 2021) como corpora de evaluación de habla típica. Las principales características de estos corpus se pueden ver en la Tabla 1.

3.1 Corpus de habla Down

El corpus PRAUTOCAL (Escudero-Mancebo et al., 2022) es un corpus de hablantes de español con SD del norte/centro peninsular que permite el análisis de aspectos específicos del habla de las personas con SD. También incluye grabaciones comparables de usuarios con desarrollo típico (DT) que sirven de referencia. El corpus se ha construido grabando interacciones con un videojuego para entrenar las competencias orales de las personas con síndrome de Down. El corpus se recopiló en seis campañas de grabación y contiene 90 locutores, con 4175 ficheros de audio distribuidos en 40 actividades diferentes asociadas al desarrollo de un videojuego educativo supervisado por terapeutas.

Para garantizar una representación homogénea del texto tanto en las referencias como en las predicciones de los modelos, ha sido necesario realizar un pre-procesamiento del corpus PRAUTOCAL Down para obtener una representación común de las transcripciones.

El etiquetado del corpus PRAUTOCAL Down contiene ciertas marcas indicativas de disfluencias que deben ser eliminadas para obtener una referencia limpia y precisa. Las marcas que nos podemos encontrar en las transcripciones de las frases del corpus son:

- Términos que han sido marcados entre los símbolos < y > indican que se ha producido una disfluencia.
- Palabras precedidas por el símbolo #, lo cual indica que dicha palabra es un *filler*, es decir, una palabra de relleno.
- Signos de puntuación (puntos y comas) que pierden su significado ortográfico y pasan a referirse a pausas que el hablante ha hecho durante la locución.

Corpus	Tipo Habla	Nº Hablantes	Nº Horas	Procedencia
FLEURS	Típica	3	12	Art. de Wikipedia
VoxPopuli	Típica	305	166	I. Parlamentarias
PRAUTOCAL Típico	Típica	30	2	Int. con Videojuego
PRAUTOCAL Down	Anómala	42	2	Int. con Videojuego

Tabla 1: Tabla resumen de las principales características de cada corpus utilizado. Los datos de todos los corpus son de la partición en español.

Además de los anteriormente citados, también se ha eliminado cualquier otro signo de puntuación, todas las tildes (debido a que el corpus no estaba correctamente etiquetado en ese sentido) y todos los signos ortográficos, como guiones, corchetes, ..., que pudiesen aparecer. Esta misma normalización ha sido también aplicada a las predicciones de *whisper*, en la evaluación con este corpus, para poder realizar una comparación justa y precisa.

Se han eliminado 67 ficheros de audio del corpus PRAUTOCAL Down que se determinó que contienen pseudo-habla (ruidos sin sentido).

Por último, se ha decidido añadir un segundo de silencio al comienzo y al de cada fichero de audio del corpus, lo que se comprobó que mejoraba el rendimiento de los reconocedores en esas frases, como ya se detectó en el trabajo de (Prananta et al., 2022).

3.2 Corpus de habla típica

En lo que respecta a los corpus FLEURS y VoxPopuli, simplemente se les ha aplicado el pre-procesamiento básico que el propio modelo *whisper* proporciona. Este procesamiento esta enfocado al trabajo plurilingüe y se basa en eliminar todo tipo de signos ortográficos y tildes, manteniendo la *ñ*.

Para el caso especial de PRAUTOCAL Típico, se le ha aplicado la misma normalización que a PRAUTOCAL Down, debido a que, aunque carece de anotaciones, incluye numerosos errores en la ubicación de los signos de puntuación (interrogaciones y exclamaciones) y de las tildes.

3.3 Eliminación de frases en bucle

Como bien se cita en Radford et al. (2023), el modelo *whisper* tiene un defecto que hace que, en ocasiones, el proceso de predicción entre en un bucle infinito (que solo termina al alcanzar el número máximo de caracteres permitidos en la generación) donde solo escribe una y otra vez la misma palabra, letra o conjunto de palabras. El propio artículo documenta que se espera que dicho comportamiento desaparezca al hacer un *fine-tuning* del modelo. En toda la experimentación que hemos realizado nos hemos topado con este problema, tanto con el modelo base, como en la gran mayoría de los *fine-tuning* que hemos realizado. El problema con estas frases es que normalmente son muy influyentes en el WER del modelo, debido a que generan una gran cantidad de inserciones que disparan su número de errores, y por tanto, el WER general. En resumen, se ha procedido a eliminar las que denominaremos "frases en bucle".

Para cada experimento de los realizados se ha creado un nuevo corpus denominado PRAUTOCAL Down Clean, a partir del PRAUTOCAL Down, pero que no incluya los ficheros de audio que generan predicciones en bucle, que pueden ser distintos para cada modelo generado. Mantener dos versiones distintas del corpus PRAUTOCAL Down por modelo, y no del resto de corpora, se debe a que este fenómeno se ve muy intensificado en este corpus, generando variaciones de WER de hasta un 20 %. En el resto de corpora el fenómeno sigue apareciendo pero en menor medida, por eso se ha decidido experimentar sólo con una versión en estos casos, que no contiene las "frases en bucle" para cada modelo.

El fenómeno de que un modelo *whisper* repita indefinidamente palabras o fragmentos de texto ante determinadas entradas de voz puede atribuirse a varios factores: limitaciones del modelo, complejidad del contexto, problemas de entrenamiento o bucles de atención. En última instancia, el fenómeno de repetición indefinida parece provenir de la complejidad inherente de la generación de lenguaje natural y las limitaciones actuales de los modelos de inteligencia artificial en este ámbito.

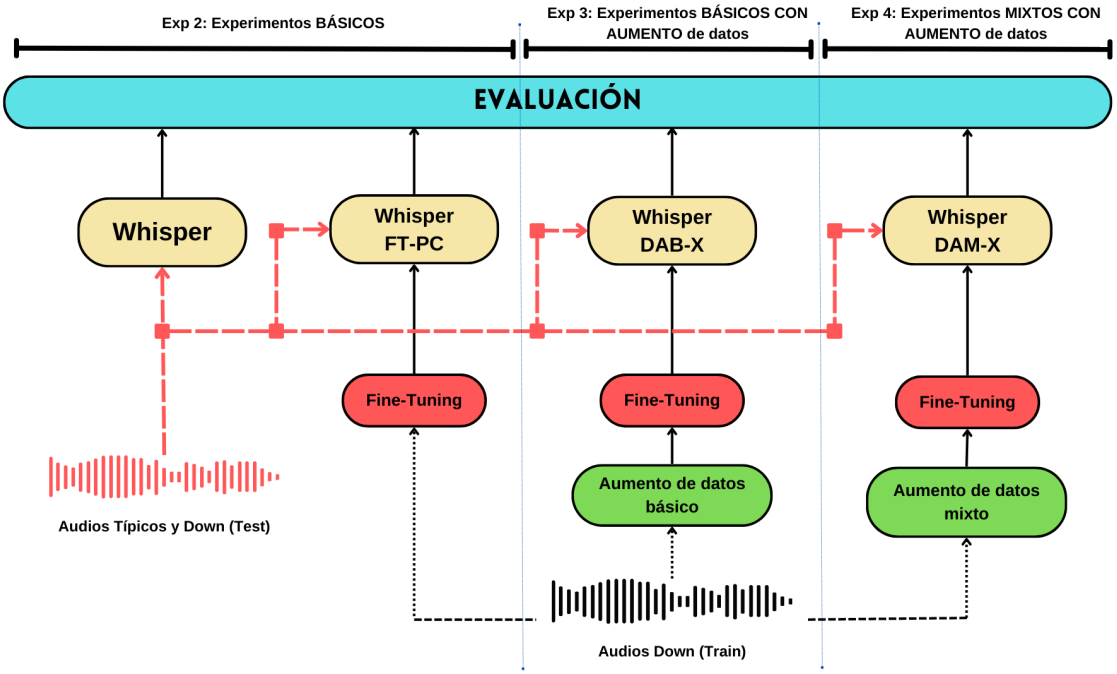

Figura 1: Diagrama explicativo de la metodología seguida en el artículo.

4 *Metodología*

En este artículo se va a usar el modelo *whisper* para realizar la experimentación, empleando el modelo de tamaño *large*. La elección de dicho modelo se debe a los buenos resultados que ofrece para español de habla típica, ya que es un modelo que ya se ha contrastado en varias ocasiones en tareas de habla anómala (Cibrian et al., 2024).

El flujo conceptual de los experimentos realizados puede verse en la Figura 1.

4.1 Adaptación de modelos

Previo a la realización de los diversos *fine-tuning*, se ha hecho un ajuste de los hiper-parámetros. Para ello, se ha utilizado el corpus FLEURS, ya que el corpus de habla Down del que se dispone es demasiado pequeño como para ser empleado en esta tarea y se podrían propiciar situaciones de *overfitting*. La búsqueda de hiper-parámetros se ha realizado siguiendo la metodología estándar: establecer un valor para cada hiper-parámetro a optimizar, posteriormente realizar un *fine-tuning* con dichos valores y con la partición *train* y evaluar el resultado con la partición *validation*. Para cada parámetro se han probado valores dentro de un intervalo inicialmente muy amplio que posteriormente se ha ido estrechando hasta alcanzar los valores óptimos: *learning rate: 2.155e-5*, *weight decay: 3.168e-4*, *epochs: 1*, usando siempre formato de coma flotante de precisión alta. La métrica utilizada para decidir que valores de los hiper-parámetros eran mejores ha sido la métrica WER.

4.1.1 Partición de datos

Debido principalmente al tamaño reducido del corpus PRAUTOCAL Down, se ha optado por un esquema de partición 60/40: 60 % de los ficheros de audio para *train* y 40 % para *test*, obviando la partición de validación. El porcentaje de muestras de *train* se ha elegido teniendo en cuenta que se aplicarán técnicas de aumento de datos sobre el corpus que harán que el conjunto de entrenamiento crezca considerablemente, lo que podría generar que la partición *test* tuviera un tamaño muy reducido en comparación a la de *train* si se optase por otros esquemas más tradicionales. Siguiendo el esquema anterior de partición, se han realizado dos divisiones diferentes, que se describen a continuación.

La **división por actividad** consiste en separar en *train* y *test* las diferentes actividades que contiene el corpus PRAUTOCAL. Todas las locuciones del 60 % de las 40 actividades diferentes que contiene el corpus irán al conjunto *train* y el 40 % restante al de *test*. Entendemos por actividad un tipo de frase pronunciada por cualquier locutor. Para cada actividad y locutor existen diferentes locu-

ciones. Realizando esta división conseguimos que ninguna partición tenga frases en común. De esta manera, una vez realizado el *fine-tuning*, al evaluar el modelo con el conjunto *test*, se enfrentará a frases nunca vistas, lo que nos permitirá observar cuál es su verdadero aprendizaje de las características propias del habla Down, al ser un modelo independiente de locutor y de tarea.

La **división aleatoria** consiste en distribuir directamente las locuciones entre *train* y *test* siguiendo los porcentajes 60-40 %, sin tener en cuenta la actividad. Con esta división es muy probable que algunas frases coincidan en ambos conjuntos, y por tanto, es posible que el modelo no solo aprenda las características intrínsecas del habla Down, sino también la pronunciación de esas frases particulares, resultando sólo independiente de locutor pero no de tarea.

4.1.2 Aumento de datos

La escasez de datos de habla anómala es un problema con el que hay que lidiar diariamente en la construcción de modelos de ASR. Con el fin de combatir esta limitación, una de las técnicas mas utilizadas es el aumento datos (Hermann y Magimai.-Doss, 2023), mediante técnicas como conversión de voz, modificación de audio, generación de voz vía TTS, ... Este problema también se manifiesta en nuestros experimentos y por eso vamos a aplicar técnicas de modificación de audio básicas sobre los ficheros de audio de PRAUTOCAL Down, como una primera aproximación al aumento de datos de habla Down.

- **Variación de la velocidad**: Esta técnica consiste en variar la velocidad del audio original, obteniendo así un nuevo audio un poco distinto. Conseguimos así simular cambios en el ritmo y en la velocidad del habla, generando de esta forma nuevos "pseudo-hablantes". Estos nuevos "pseudo-hablantes" pueden ayudar al modelo a generalizar mejor para nuevos ritmos y velocidades a los cuales el modelo no esta acostumbrado. La modificación no debe ser muy grande, dado que sino el modelo estaría aprendiendo ritmos y velocidades irreales con los cuales nunca se va a encontrar. Es por ello que la velocidad del audio se variará a 0.9 y 1.1 empleando la biblioteca *wave*.[2]
- **Introducción de ruido**: Esta técnica consiste en añadir ruido de diferente naturaleza al audio original. Se han considerado 2 tipos de ruido. El primero es el ruido blanco cuya principal característica es que tiene una densidad espectral de potencia constante. Esto significa que el ruido contiene todas las frecuencias y todas ellas tienen la misma potencia. El segundo tipo es el ruido de color, el cual no tiene una densidad espectral constante, sino que dependiendo de la forma que tenga la onda se clasificará como ruido de un color o de otro. Esta nueva modificación puede ayudar al modelo a trabajar con ficheros de audio cuya calidad no sea del todo perfecta o que contengan sonidos de fondo que puedan distorsionar el rendimiento del modelo. Dado la gran cantidad de tipos de ruidos que existen se ha decidido utilizar solo dos tipos, el ruido blanco y el ruido rosa. La elección del primero se debe a sus características únicas explicadas anteriormente. En cuanto al segundo, se ha decidido elegir el ruido de color rosa debido a que es el más frecuente en experimentos de este estilo. Para generar los ruidos de color rosa se ha utilizado el algoritmo propuesto en Timmer y Koenig (1995).
- **Variación del tono**: Esta técnica consiste en variar el tono (*pitch*) del audio original obteniendo así nuevos ficheros de audio ligeramente distintos a los originales. La aplicación de esta técnica genera nuevos "pseudo-hablantes" que pueden facilitar la generalización al modelo. En la implementación de esta técnica se va a subir/bajar un cierto número de semitonos al audio original. La variación no debe ser muy grande, por lo que las modificaciones que se van a realizar van a consistir en subir/bajar 2 semitonos. Se ha utilizado la función *pitch_shift()* de la biblioteca *librosa*.[3]

Las técnicas descritas se van a aplicar sobre el corpus PRAUTOCAL Down, tanto de una en una como en grupos de dos, tres y cuatro, con el objetivo de aumentar los pocos datos de los que se disponen. Al incrementar el número de ficheros de audio se espera obtener una mejoría en el rendimiento del modelo, debido a que al aumentar la cantidad de

[2]https://docs.python.org/3/library/wave.html

[3]https://librosa.org/doc/latest/index.html

datos, el modelo dispone de más muestras para aprender las características intrínsecas de las mismas, lo que facilita una generalización más efectiva.

El procedimiento de aumento de datos consistirá en generar un nuevo audio por cada instancia de cada combinación de técnicas aplicada. Así, si aplicamos la técnica de *Variación de la Velocidad* sobre el corpus, triplicaremos su tamaño, ya que tendremos los ficheros de audio originales, los ficheros de audio modificados a velocidad 0.9 y los modificados a velocidad 1.1. Cabe aclarar que el aumento de datos se aplica sólo sobre la partición *train*.

4.2 Modelos evaluados

El objetivo de este trabajo es desarrollar un modelo de ASR que mejore el rendimiento de los sistemas para personas con SD. Se llevan a cabo cuatro tipos de experimentos distintos, cada uno diseñado para evaluar el rendimiento del modelo en diferentes escenarios. Estos experimentos incluyen el uso de corpus aumentados para habla Down, así como el análisis del nuevo modelo *whisper-large-v3*. Se realizan ajustes de hiperparámetros con el corpus FLEURS y se llevan a cabo experimentos de fine-tuning.

En nuestro estudio, no sólo nos interesa ver el rendimiento del modelo para habla Down, sino que también nos interesa ver cómo evoluciona el rendimiento para habla típica en función del modelo preparado. Es por eso que todo experimento realizado será evaluado tanto para habla Down como para habla típica.

Se han realizado cuatro tipos de experimentos diferentes:

- **Experimento Base**: Se evalúa el modelo base (BASE) con diferentes corpus y se realiza un primer fine-tuning (FT-PC) con cada división del corpus PRAUTOCAL Down.
- **Experimento Base con aumento de datos**: Se realizan 4 experimentos distintos por división, utilizando corpus aumentados con una sola técnica de las descritas (DAB-X). Se busca determinar la utilidad de las técnicas de aumento de datos para ayudar al modelo a generalizar.
- **Experimentos Mixtos con aumento de datos**: Se llevan a cabo 11 experimentos distintos para cada tipo de división, combinando dos, tres o cuatro de las técnicas de aumento de datos anteriormente descritas, cada una de las cuales genera el correspondiente fichero de audio extra.
- **Experimentos con *whisper-large-v3***: Se replican 3 experimentos sobre el nuevo modelo *whisper-large-v3*, seleccionando los más representativos. Se ajustan hiperparámetros con el corpus FLEURS.

4.3 Métricas

Además de la métrica estándar WER, aplicada tanto a frases sueltas como a colecciones de frases de un mismo usuario o de una misma partición, se ha empleado el **Análisis con BERT F1 Score**.

BERTScore (Zhang et al., 2020) es una métrica de evaluación automática para la generación de texto que calcula una puntuación de similitud para cada token en la oración candidata con cada token en la oración de referencia. Aprovecha las incorporaciones contextuales previamente entrenadas de los modelos BERT y relaciona palabras en oraciones candidatas y de referencia mediante similitud de coseno.

Analizar el rendimiento de un modelo en base al número de palabras coincidentes entre la referencia y la predicción puede llegar a ser algo pobre. Pueden existir casos con WER alto pero debido solamente a fallos puntuales en ciertas palabras que no son del todo importantes para comprender el significado de la frase. Para solucionar este problema usamos el BERT F1 Score, que ha sido ya utilizada para evaluar el rendimiento de sistemas ASR (Tobin et al., 2022). Esta métrica no se fija en la similitud ortográfica de las frases, sino en su similitud semántica. Consideramos que el análisis con esta métrica, junto con los resultados expuestos anteriormente, ayudará a dar unos resultados más completos.

De todos los experimentos realizados solo se va a aplicar esta métrica con los siguientes experimentos (al igual que en la experimentación con *whisper-large-v3*): **BASE**, **FT-PC** y **DAB-2**. La métrica se va a aplicar para las particiones *test* de los corpus que se han venido utilizando, y para ambas versiones de *whisper* (v2 y v3). Los resultados se pueden ver en la Tabla 4.

Corpus	BASE	FT-PC	DAB-1	DAB-2	DAB-3	DAB-4
División Aleatoria						
VoxPopuli (ES)	8.422	15.585	15.390	16.083	16.940	17.602
PRAUTOCAL Típico (ES)	3.409	2.193	1.929	1.982	2.167	2.511
PRAUTOCAL Down (SD)	64.315	23.565	14.653	14.576	22.559	22.463
PRAUTOCAL Down Clean (SD)	35.858	17.354	14.653	**14.576**	15.747	15.021
Nº ficheros de audio (Train)	-	1208	2416	2416	3624	3624
División por Actividades						
VoxPopuli (ES)	8.422	15.941	16.662	15.409	19.100	18.742
PRAUTOCAL Típico (ES)	4.239	7.781	6.815	9.149	10.789	10.572
PRAUTOCAL Down (SD)	56.166	44.645	29.091	27.283	28.164	29.671
PRAUTOCAL Down Clean (SD)	37.668	28.376	27.355	**27.029**	28.164	29.551
Nº ficheros de audio (Train)	-	1275	2550	2550	3825	3825

Tabla 2: WER (%) para cada corpus y experimento realizado con *whisper-large-v2*. BASE: sin fine-tuning. FT-PC: Fine-tuning con PRAUTOCAL Down. DAB-X: Fine-tuning con PRAUTOCAL Down, pero aplicándole una técnica de aumento de datos (1- Ruido Blanco, 2- Ruido Rosa, 3- Variación Velocidad, 4- Variación Tono). El conjunto test contiene: 875 para la división aleatoria y 808 para la división por actividades.

Corpus	BASE	FT-PC	DAB-2
División Aleatoria			
VoxPopuli (ES)	11.214	13.116	13.657
PRAUTOCAL Típico (ES)	3.013	1.929	2.748
PRAUTOCAL Down (SD)	48.734	24.995	22.695
PRAUTOCAL Down Clean (SD)	32.232	15.908	**15.471**
División por Actividades			
VoxPopuli (ES)	11.214	12.714	12.726
PRAUTOCAL Típico (ES)	3.488	7.137	7.352
PRAUTOCAL Down (SD)	52.851	51.553	32.221
PRAUTOCAL Down Clean (SD)	32.919	25.336	**25.175**

Tabla 3: WER (%) para cada corpus y experimento realizado con *whisper-large-v3*. BASE: sin fine-tuning. FT-PC: Fine-tuning con PRAUTOCAL Down. DAB-2: Fine-tuning con PRAUTOCAL Down, pero aplicándole la técnica que mejores resultados ha dado en la experimentación con *whisper-large-v2* (Ruido Rosa).

5 Resultados

Como muestran las Tablas 2 y 4, el modelo base permite alcanzar un resultado de WER del 8.4 % y un BERT Score F1 de 0.97 para el corpus VoxPopuli (habla típica español). Los resultados mejoran cuando este modelo base se aplica al corpus PRAUTOCAL Típico (WER: 3.4 % y BERT:0.988) y empeoran mucho cuando se aplica al corpus PRAUTOCAL Down (WER:> 35,9 % y BERT:< 0,9).

La aplicación de *fine-tuning* hace que los resultados empeoren para VOXPOPULI entre un 8 % y un 9 %y mejoren en el caso del corpus PRAUTOCAL. Está mejora es especialmente relevante en el caso de habla Down, donde las tasas pasan de 35.858 % a un 17.354 % para el caso de eliminación de frases en bucle y división aleatoria.

Las técnicas de aumento de datos (DAB) aportan una mejora que es muy poco significativa en el dataset Clean con división por actividades (en torno a un 1 %) y de hasta 9 % para dataset sin eliminar frases en bucle y división aleatoria. Esto claramente nos indica que el modelo *whisper*, en el caso de la división aleatoria, esta aprendiendo frases en el proceso de *fine-tuning* que luego aparecen en la partición *test*. Al aumentar el tamaño del corpus, se incrementa el número de repeticiones de cada frase, facilitando así el aprendizaje de la pronunciación de frases es-

Modelo	Corpus	BASE	FT-PC	DAB-2
V2	**División Aleatoria**			
	VoxPopuli (ES)	0.975	0.953	0.951
	PRAUTOCAL Típico (ES)	0.988	0.991	0.992
	PRAUTOCAL Down Clean (SD)	0.882	0.947	0.953
	División por Actividades			
	VoxPopuli (ES)	0.975	0.950	0.951
	PRAUTOCAL Típico (ES)	0.984	0.976	0.973
	PRAUTOCAL Down Clean (SD)	0.876	0.914	0.916
V3	**División Aleatoria**			
	VoxPopuli (ES)	0.965	0.961	0.959
	PRAUTOCAL Típico (ES)	0.989	0.992	0.989
	PRAUTOCAL Down Clean (SD)	0.893	0.948	0.952
	División por Actividades			
	VoxPopuli (ES)	0.965	0.959	0.961
	PRAUTOCAL Típico (ES)	0.987	0.976	0.976
	PRAUTOCAL Down Clean (SD)	0.895	0.925	0.925

Tabla 4: Análisis con la métrica BERTScore F1 de los modelos más significativos que se han generado a lo largo de toda la experimentación.

pecificas. El escaso incremento obtenido en la división por actividades, permite concluir que sólo con el *fine-tuning* **FT-PC** ya nos acercamos mucho al máximo aprendizaje que se puede obtener del corpus PRAUTOCAL.

Si nos restringimos al modelo *whisper* v2, del 18.5 % de mejora en el resultado de WER que se obtiene con la división aleatoria entre **FT-PC** y **BASE**, la mitad aproximadamente (9.3 %) es atribuible a mejora de resultado WER en reconocimiento de habla Down, que es la diferencia entre el WER de **BASE** y **FT-PC** en la división por actividades.

En cuanto a los resultados con cada una de las distintas técnicas de aumento de datos, podemos ver que en el caso de la división aleatoria todas mejoran, por lo tanto, podemos concluir que existe una cierta mejora que viene directamente proporcionada por el aumento de datos y no por la técnica empleada (aproximadamente un 2 %). En cuanto a la división por actividad podemos comprobar que las técnicas de *Variación de Velocidad* y *Variación de Ruido* empeoran el rendimiento del modelo. Esto se puede deber a que la simulación de "pseudo-hablantes" no sea del todo realista, y por tanto, estén introduciendo en el corpus características no reales. Los mejores resultados, en ambas divisiones, se obtienen con la técnica de *Introducción de Ruido Rosa*. Esta mejora puede venir explicada por la naturaleza propia de las personas con SD. Estas personas cometen una gran cantidad de bloqueos, repeticiones de palabras, elongaciones y ruidos entre palabras, que aparecen cuando sufren complicaciones en la frase que desean pronunciar. Estos sonido producidos pueden parecerse al ruido que se introduce con esta técnica, y por tanto, conseguir que el modelo se adapte mejor.

En cuanto a los experimentos mixtos con aumento de datos, ninguno de los experimentos obtuvo una mejora en el rendimiento del modelo. Los resultados de estos experimentos no se incluyen en el artículo por cuestiones de espacio.

La utilización de la versión v3 del modelo *whisper*, publicado mientras se preparaba este trabajo, no conlleva apenas diferencias de comportamiento frente a la versión v2 ya comentada.

Por lo que respecta a la métrica BERTScore F1, podemos ver que, en líneas generales, el modelo *whisper* (v2 y v3) genera predicciones de texto a partir de voz que tienen una muy buena correspondencia semántica incluso cuando el WER pueda no ser bajo, tanto en habla típica como Down. Así, el peor resultado obtenido es *0.876* – tengamos en cuenta que la métrica BERTScore F1 se mueve en un rango 0(pésimo)-1(óptimo). La WER más baja, obtenida para división aleatoria y **DAB-2**, es de 14.576 %, y se corresponde con

una BERTScore F1 de 0.953.

En las Tablas 2 a 4 se puede ver que existe una correlación casi total entre los resultados de WER y de BERT F1 Score para los diferentes modelos aplicados a cada corpus analizado (r-pearson: -0.998, p-value=4.1e-20): cuando el WER mejora (baja) el BERT también (sube), y viceversa. Aun así, se puede apreciar cómo grandes cambios en el WER no tienen porqué conllevar grandes variaciones en el BERT F1 Score (por ejemplo, el salto entre el modelo BASE y el FT-PC de *whisper-large-v2* evaluado con VoxPopuli), y cómo existen casos en los que pequeñas variaciones del WER suponen cambios más notables en el BERT F1 Score (como la diferencia entre FT-PC y DAB-2 de *whisper-large-v2* evaluado en la división Aleatoria de PRAUTOCAL Down). Esta variabilidad reside en el valor semántico de las palabras erradas/acertadas por *whisper*, haciendo así que no haya una relación total entre el WER y el BERTScore, y dando así valor a los resultados obtenidos con esta métrica.

En cuanto al habla Down, obtenemos una prueba más de lo influyente que es que existan frases comunes en *train* y *test*. Los resultados muestran que en los experimentos que involucran un *fine-tuning* con PRAUTOCAL Down el resultado de la división aleatoria es siempre mejor que el de la división por actividad. Esto se debe en su totalidad a la distribución de la división.

Por último, podemos corroborar que la aplicación del *fine-tuning* con PRAUTOCAL Down es muy beneficiosa para mejorar tanto el acierto en la predicción de palabras del modelo (WER), como su comprensión semántica. En el caso del aumento de datos aplicados a la división por actividad, se ve como la diferencia es prácticamente inexistente, siguiendo la tendencia que mostraba el WER, y demostrando así que el uso de ampliación de datos no es extremadamente beneficioso para que el modelo aprenda más sobre el habla Down. Por otro lado, sí que es ciertamente beneficioso en el caso de la división aleatoria, siendo esto debido a la naturaleza de la distribución originada por esta división.

6 Conclusiones y trabajo futuro

Los resultados presentados en este trabajo muestran que los modelos base *whisper-large-v2* y *whisper-large-v3* tienen un rendimiento muy malo en el reconocimiento de habla Down española. Aun así cabe destacar que su desempeño en esta tarea, o similares, es superior a la gran mayoría de modelos ASR actuales (Cibrian et al., 2024). Por otro lado, se ha demostrado que aplicar la técnica de *fine-tuning* con un corpus Down sobre estos modelos, proporciona grandes mejoras, tanto en la capacidad de adaptación del modelo al habla Down (división por actividad), como en la adaptación específica a la tarea del corpus PRAUTOCAL (división aleatoria). La mejora anteriormente citada, siempre lleva consigo un decremento en el rendimiento del modelo para habla típica, por lo que podemos concluir que el *fine-tuning* del modelo *whisper* tiene una capacidad de aprendizaje limitada. En lo que respecta a la aplicación de técnicas de aumento de datos, se observa que su eficacia es limitada cuando nos centramos en obtener mejoras en características generales del habla Down, pero sí resultan útiles para mejorar el rendimiento en escenarios dependientes de tarea, en el corpus en cuestión. Finalmente, en relación a la experimentación realizada, podemos concluir que la versión v3 del modelo *whisper* no es esencialmente mejor que la versión v2, en lo que a reconocimiento de habla Down española respecta.

Como trabajo futuro queda la exploración de técnicas de aumento de datos más complejas, como puede ser conversión de voz, o la aplicación de un TTS para simular la voz Down y aumentar la variedad léxica del corpus PRAUTOCAL. También queda como trabajo pendiente realizar un estudio más en profundidad de los resultados obtenidos en este artículo, centrando la atención en explicar porque se han obtenido estos resultados y su relación con las características propias de los hablantes Down. Además, como trabajo futuro, planteamos analizar más a fondo el fenómeno por el que en ocasiones el proceso de predicción entra en un bucle infinito. Finalmente, queda pendiente realizar un estudio y una experimentación más profunda y completa sobre el nuevo modelo *whisper-large-v3*.

Agradecimientos

Este trabajo ha sido realizado en el marco del proyecto **PID2021-126315OB-I00** que ha sido financiado por **MCIN / AEI / 10.13039/501100011033 / FEDER, EU**.

Bibliografía

Almadhor, A., R. Irfan, J. Gao, N. Saleem, H. Tayyab Rauf, y S. Kadry. 2023. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications*, 222:119797.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. American Psychiatric Publishing, Arlington, VA.

Bhat, C. y H. Strik. 2020. Automatic assessment of sentence-level dysarthria intelligibility using blstm. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):322–330.

Caton, S. y M. Chapman. 2016. The use of social media and people with intellectual disability: A systematic review and thematic analysis. *Journal of intellectual and developmental disability*, 41(2):125–139.

Chapman, R. S. 1997. Language development in children and adolescents with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, 3(4):307–312.

Cibrian, F. L., K. Anderson, C. M. Abrahamsson, V. G. Motti, y others. 2024. Limitations in speech recognition for young adults with Down syndrome. *Research Square (Preprint Version 1)*.

Cleland, J., S. Wood, W. Hardcastle, J. Wishart, y C. Timmins. 2010. Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. *International journal of language & communication disorders*, 45(1):83–95.

Conneau, A., M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, y A. Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. En *2022 IEEE Spoken Language Technology Workshop (SLT)*, páginas 798–805.

De Russis, L. y F. Corno. 2019. On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, 5:163–172.

Escudero-Mancebo, D., M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, y V. Flores-Lucas. 2022. PRAUTOCAL corpus: a corpus for the study of Down syndrome prosodic aspects. *Language Resources and Evaluation*, 56:191–224, Mayo.

Feng, J., J. Lazar, L. Kumin, y A. Ozok. 2010. Computer usage by children with Down syndrome: Challenges and future research. *ACM Transactions on Accessible Computing (TACCESS)*, 2(3):1–44.

Green, J. R., R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, y K. Tomanek. 2021. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. En *Proc. Interspeech 2021*, páginas 4778–4782.

Hermann, E. y M. Magimai.-Doss. 2023. Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation. En *Proc. INTERSPEECH 2023*, páginas 156–160.

Hu, R., J. Feng, J. Lazar, y L. Kumin. 2013. Investigating input technologies for children and young adults with Down syndrome. *Universal access in the information society*, 12:89–104.

Janbakhshi, P., I. Kodrasi, y H. Bourlard. 2021. Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks. En *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 7328–7332. IEEE.

Jiao, Y., M. Tu, V. Berisha, y J. Liss. 2018. Simulating dysarthric speech for training data augmentation in clinical speech applications. En *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, páginas 6009–6013. IEEE.

Kitzing, P., A. Maier, y V. L. Åhlander. 2009. Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, 34(2):91–96.

Kumin, L. 2012. *Early communication skills for children with Down syndrome: A guide for parents and professionals*. Woodbine House, 3ª edición.

Laws, G. y D. V. Bishop. 2004. Verbal deficits in Down's syndrome and specific language impairment: a comparison. *International Journal of Language & Communication Disorders*, 39(4):423–451.

Lea, C., Z. Huang, J. Narain, L. Tooley, D. Yee, D. T. Tran, P. Georgiou, J. P. Bigham, y L. Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. En *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, páginas 1–16.

MacDonald, R. L., P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, J. R. Green, y K. Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. En *Interspeech 2021*, páginas 4833–4837.

Martin, G. E., J. Klusek, B. Estigarribia, y J. E. Roberts. 2009. Language characteristics of individuals with Down syndrome. *Topics in language disorders*, 29(2):112–132.

Mitra, V., Z. Huang, C. Lea, L. Tooley, S. Wu, D. Botten, A. Palekar, S. Thelapurath, P. Georgiou, S. Kajarekar, y J. Bigham. 2021. Analysis and Tuning of a Voice Assistant System for Dysfluent Speech. En *Proc. Interspeech 2021*, páginas 4848–4852.

Prananta, L., B. Halpern, S. Feng, y O. Scharenborg. 2022. The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition. En *Proc. Interspeech 2022*, páginas 36–40.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever. 2023. Robust speech recognition via large-scale weak supervision. En *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Rosen, K. y S. Yampolsky. 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16(1):48–60.

Schultz, B. G., V. S. A. Tarigoppula, G. Noffs, S. Rojas, A. van der Walt, D. B. Grayden, y A. P. Vogel. 2021. Automatic speech recognition in neurodegenerative disease. *International Journal of Speech Technology*, 24(3):771–779.

Shahamiri, S. R. 2021. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:852–861.

Shor, J., D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, y Y. Matias. 2019. Personalizing asr for dysarthric and accented speech with limited data. En *Interspeech 2019*, interspeech2019. ISCA, Septiembre.

Tanis, E. S., S. Palmer, M. Wehmeyer, D. K. Davies, S. E. Stock, K. Lobb, y B. Bishop. 2012. Self-report computer-based survey of technology use by people with intellectual and developmental disabilities. *Intellectual and developmental disabilities*, 50(1):53–68.

Timmer, J. y M. Koenig. 1995. On generating power law noise. *Astronomy and Astrophysics, v. 300, p. 707*, 300:707.

Tobin, J., Q. Li, S. Venugopalan, K. Seaver, R. Cave, y K. Tomanek. 2022. Assessing ASR Model Quality on Disordered Speech using BERTScore. En *Proc. 1st Workshop on Speech for Social Good (S4SG)*, páginas 26–30.

Tobin, J. y K. Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. En *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 6637–6641.

Tomanek, K., F. Beaufays, J. Cattiau, A. Chandorkar, y K. C. Sim. 2021. On-device personalization of automatic speech recognition models for disordered speech. *arXiv:2106.10259*.

Venugopalan, S., J. Shor, M. Plakal, J. Tobin, K. Tomanek, J. R. Green, y M. P. Brenner. 2021. Comparing Supervised Models and Learned Speech Representations for Classifying Intelligibility of Disordered Speech on Selected Phrases. En *Interspeech 2021*, páginas 4843–4847.

Wang, C., M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, y E. Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. En C. Zong F. Xia W. Li, y R. Navigli, editores, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 993–1003, Online, Agosto. Association for Computational Linguistics.

Wong, B., C. Brebner, P. McCormack, y A. Butcher. 2015. Word production inconsistency of Singaporean-English-speaking adolescents with Down Syndrome. *International journal of language & communication disorders*, 50(5):629–645.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, y Y. Artzi. 2020. Bertscore: Evaluating text generation with bert. En *International Conference on Learning Representations*.