UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Modelado de embeddings para la reducción de sesgo en sistemas de reconocimiento de locutor

Máster Universitario en Ingeniería de Telecomunicación

Autor: Aguilera Sepúlveda, Almudena
Tutor: Lozano Díez, Alicia
Ponente: González Rodríguez, Joaquín

FECHA: Junio, 2024

Modelado de embeddings para la reducción de sesgo en sistemas de reconocimiento de locutor

AUTOR: Aguilera Sepúlveda, Almudena
TUTOR: Lozano Díez, Alicia
PONENTE: González Rodríguez, Joaquín

Audio, Data Intelligence and Speech - AUDIAS
Dpto. Tecnología Electrónica y de las Comunicaciones, EPS
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2024

But courage need not be remembered… for it is never forgotten.
- Princess Zelda, from Breath Of The Wild

A una de las personas que más admiro y quiero,
que siempre ha estado acompañándome,
mi abuela

Agradecimientos

A mi familia y amigos, que habéis estado de manera incansable durante esta etapa, apoyándome incondicionalmente, y aguantando tanto mis buenos momentos como los malos, los momentos de hiperactividad en los que parecía un torbellino con patas y aquellos en los que parecía que me había rendido. Todo vuestro amor, cariño y compañía han sido fundamentales durante este viaje.

A mi tutora, Alicia Lozano Díez, por confiar para que continuase este trabajo, y, toda la ayuda y paciencia que ha tenido conmigo. También a todos los profesores con los que he tenido la suerte de encontrarme y aprender durante este proceso, no solo por enseñarme, sino también por hacer que me supere a mí misma cada día, no habría podido llegado hasta aquí sin todo esto.

Y en especial quiero terminar agradeciendo a una de las personas más maravillosas que conozco, a mi abuela. Por ser ese ejemplo de resiliencia y valentía frente a todos los desafíos. Demostrándome que da igual que tengas 24 o 92 años, siempre se puede aprender algo nuevo y mejorar. Por ser esa segunda madre que no sabía que necesitaba, una fuente inacabable de amor, cariño y comprensión. Podría decir mil cosas más y aun no habría acabado… En resumen, por ser mi mayor ejemplo a seguir y ese faro que me guía cuando me siento perdida. A ella, a ti si lo estas leyendo, nunca podré terminar de agradecerte lo que has hecho por mí, por eso te dedico este trabajo, porque si he llegado hasta aquí, si me he convertido en lo que soy hoy en día, ha sido un 10% la cabezonería que me viene de serie y un 90% tú confianza en mí.

Resumen

En este trabajo de fin de master (TFM), analizamos las causas que generan las faltas de disparidad que se producen en los sistemas de verificación de entre los distintos grupos de locutores, sesgados en función de las siguientes características privadas: edad, género y acento.

Para realizar esta investigación y análisis de las causas, generamos tres experimentos en donde evaluamos la falta de disparidad entre resultados entre los grupos sesgados, y, analizamos si esta disparidad se está produciendo a causa de los datos, o, se produce por el propio sistema. Tras el análisis de los resultados obtenidos a través de estos tres experimentos, generamos una posible solución en el cuarto experimento, que busca disminuir la falta de imparcialidad observada.

En el primer experimento evaluamos uno de los modelos del estado del arte con la base de datos (BBDD) VoxCeleb, analizando los sesgos que se producen en género y acento. Para el resto de experimento, empleamos el mismo modelo de red que se usa en este modelo: ResNet34L. En el segundo y tercer experimento, empleamos el mismo modelo de red que se emplea en VoxCeleb, pero en lugar de usar uno de los modelos ya entrenamos, generamos un modelo con la base de datos (BBDD) Mozilla Common Voice: FairVoice. En el segundo experimento, empleamos la base de datos desbalanceada, mientras que, en el tercer experimento, hacemos una modificación de esta para que se encuentre balanceada, teniendo el mismo número de locutores entre los grupos de género y edad.

Con los resultados obtenidos, se han planteado las hipótesis que, en estos casos, la falta de equilibrio está produciendo que los sistemas generen sesgos entre los datos. Como una posible solución, planteamos el experimento 4: utilizar el modelo equilibrado del experimento 3 como base, para generar un modelo nuevo basado utilizando Fine-Tuning, y, entrenado y evaluando el modelo con las mismas listas que en el experimento 2.

Los resultados satisfactorios de la solución planteada en el experimento 4, nos lleva a plantear una propuesta para reducir las tasas de disparidad entre los grupos de género, edad y una mezcla entre ambos, y, planteamos la posibilidad de abrir una nueva línea de investigación enfocada en esta clase de soluciones para paliar el problema y mejorar las tasas de error.

Palabras Clave

Tasa de error (EER), género, edad, acento, sesgos, disparidad de resultado (DS), base de datos, mínimo valor de la función de coste (MinDCF), perdida de calibración (Cllr), reconocimiento de locutor, audio, voz, neutralidad, redes neuronales, verificación de locutor, locutores.

Abstract

In this thesis, we analyze the causes that generate the lack of disparity that occurs in the verification systems between the different groups of speakers, biased according to the following private characteristics: age, gender and accent.

To carry out this investigation and analysis of the causes, we generate three experiments where we evaluate the lack of disparity between results between the biased groups, and we analyze whether this disparity is occurring due to the data, or is produced by the system itself. After analyzing the results obtained through these three experiments, we generated a possible solution in the fourth experiment, which seeks to reduce the lack of impartiality observed.

In the first experiment we evaluated one of the state-of-the-art models with the VoxCeleb database (BBDD), analyzing the biases that occur in gender and accent. For the rest of the experiment, we use the same network model used in this model: ResNet34L. In the second and third experiments, we used the same network model that is used in VoxCeleb, but instead of using one of the models we already trained, we generated a model with the Mozilla Common Voice database (DB): FairVoice. In the second experiment, we use the unbalanced database, while, in the third experiment, we modify it so that it is balanced, having the same number of speakers between the gender and age groups.

With the results obtained, the hypotheses have been raised that, in these cases, the lack of balance is causing the systems to generate biases between the data. As a possible solution, we propose experiment 4: use the balanced model from experiment 3 as a base, to generate a new model based on Fine-Tuning, and, train and evaluate the model with the same lists as in experiment 2.

The satisfactory results of the solution proposed in experiment 4 lead us to propose a proposal to reduce the disparity rates between gender and age groups and a mixture between the two, and we propose the possibility of opening a new line of focused research in this kind of solutions to alleviate the problem and improve error rates.

Keywords

Error rate (EER), gender, age, accent, biases, disparity score (DS), data base, minimum value of Detection Cost Function (MinDCF), calibration los (Cllr), speaker recognition, audio, voice, fairness, neuronal network, speaker verification, speakers.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE FÓRMULAS

1 Introducción

1.1 Motivación

En los últimos años, con el rápido avance tecnológico y el aumento del uso de la inteligencia artificial, principalmente el desarrollo de las redes neuronales profundas se ha producido un aumento de los usos de sistemas de reconocimiento de locutor. Podemos ver el auge de estos sistemas en aplicaciones comerciales (a través de los asistentes por voz como: Alexa, Google Assitant, Siri, Cortana, Bixby, etc.) y en herramientas destinadas al análisis forense y aplicaciones de seguridad.

Especialmente en estos dos últimos grupos, se han incrementado la duda y preocupación sobre la siguiente cuestión: "¿dentro de estos sistemas, puede existir una falta de neutralidad en función de las características privadas de los locutores?". En los últimos años, esta cuestión ha desarrollado una línea de investigación con el fin de analizar este problema y buscar una posible solución. En las investigaciones llevadas a cabo, se ha demostrado la existencia de este problema, y en algunos de ellos se han probado distintas soluciones para intentar paliar este problema.

La principal motivación de este trabajo es evaluar los sesgos de los sistemas de reconocimiento de locutor en el estado del arte, ampliando la investigación que se inició en un trabajo de fin de grado previo **¡Error! No se encuentra el origen de la referencia.**, y, recopilando más datos a través de nuevos métodos que nos indiquen con mayor claridad no solo las posibles faltas de neutralidad, sino también determinar las posibles fuentes del problema: los datos y/o el propio sistema.

1.2 Objetivos

El objetivo principal de este trabajo es analizar los posibles patrones de sesgos que se producen en los sistemas del estado del arte, y analizar las posibles fuentes de estos. Para lograr este objetivo, vamos a realizar distintas tareas que se pueden dividir en los siguientes puntos:

- Analizar los patrones de sesgo y las posibles causas usando los modelos del estado del arte con bases de datos desbalanceadas, esto significa que no tendremos la misma cantidad de locutores en los distintos grupos.

- Con los mismos sistemas del estado de arte, analizar esos mismos patrones y causas cuando empleamos una base de datos balanceada. Estas bases balanceadas contendrán la misma cantidad de locutores en los grupos sesgados en género y edad.

- Emplear un *framework* generado en estudios recientes [2][20][26] para analizar la calibración de los datos empleados en los sistemas. En este *framework* emplean dos métodos de calibración: ad-Hoc (calibración mediante embeddings) y post-Hoc (calibración mediante los scores), para poder tener una

mayor precisión a la hora de determinar si existe falta de neutralidad o es producida por una falta de representación dentro de la base de datos.

- Por último, buscamos generar distintas soluciones que palien o al menos reduzcan el sesgo generado en los sistemas del estado del arte, como complemento a esta solución pretendemos conseguir una reducción de la tasa de error que encontremos en el análisis de los sistemas.

1.3 Organización de la memoria

- **Capítulo 1: Introducción**

Este primer capítulo explicamos de forma resumida la motivación para realizar este trabajo, teniendo en cuenta las investigaciones previas que se han llevado en este campo. También, detallamos los objetivos que queremos realizar a lo largo de esta investigación, y, cómo estructuraremos la memoria.

- **Capítulo 2: Estado del arte**

Este segundo capítulo se describirán las tecnologías que se han empleado a lo largo de las décadas hasta el actual estado del arte para el desarrollo de los sistemas de reconocimiento de locutor.

Explicaremos que es la neutralidad o *fairness*, aplicado a los sistemas de aprendizaje automáticos, además de cuáles son los métodos empleados para conseguir esta neutralidad, o, al menos reducir los sesgos que se generan, incluido el método de calibración que emplearemos en los experimentos.

- **Capítulo 3: Entorno experimental**

El tercer capítulo de esta memoria, lo dedicamos a explicar los distintos conjuntos de herramientas, bases de datos y métodos para poder realizar y evaluar nuestros experimentos y el comportamiento de los modelos que generamos.

Además de explicar los elementos que empleamos, profundizaremos en las características y la distribución de los locutores que encontramos en las dos bases de datos a emplear: VoxCeleb (base de datos perteneciente al estado del arte) y Mozilla Common Voice (base de datos usada en unos de los experimentos). Detallamos esta información, ya que influyen a la hora de generar los distintos modelos.

- **Capítulo 4: Experimentos, desarrollo y resultados**

Este capítulo está dedicado a explicar cada uno de los experimentos que llevamos a cabo y sus resultados. Para describir cada experimento utilizaremos una estructura común, motivación que nos ha llevado a realizarlo, los procesos previos: análisis de los datos y códigos auxiliares para la preparación del experimento, y, el

desarrollo de este. En este capítulo también mostraremos los resultados de cada uno de ellos (tablas, gráficas, etc.), y, explicaremos los resultados obtenidos.

- **Capítulo 5: Conclusiones y trabajos futuros**

Para concluir el trabajo llevado a cabo, contamos con el capítulo 5 para expresar nuestras conclusiones después del trabajo realizado, de manera concisa y resumida. A partir de estas conclusiones propondremos posibles continuaciones dentro de esta línea de investigación.

2 Estado del arte

Para entender cómo funciona un sistema de reconocimiento de locutor, primero tenemos comprender que la voz presenta una serie de características: fonológicas y prosódicas, propias de cada individuo. Por ese motivo somos capaces de distinguir cada voz y realizar una identificación de la persona que está hablando.

Los sistemas de reconocimiento se basan en técnicas que extraen esas características inherentes para llevar a cabo la identificación o verificación del sujeto, a través de muestras de su voz (audios). En la Figura 2-1 se representa estas dos tareas.

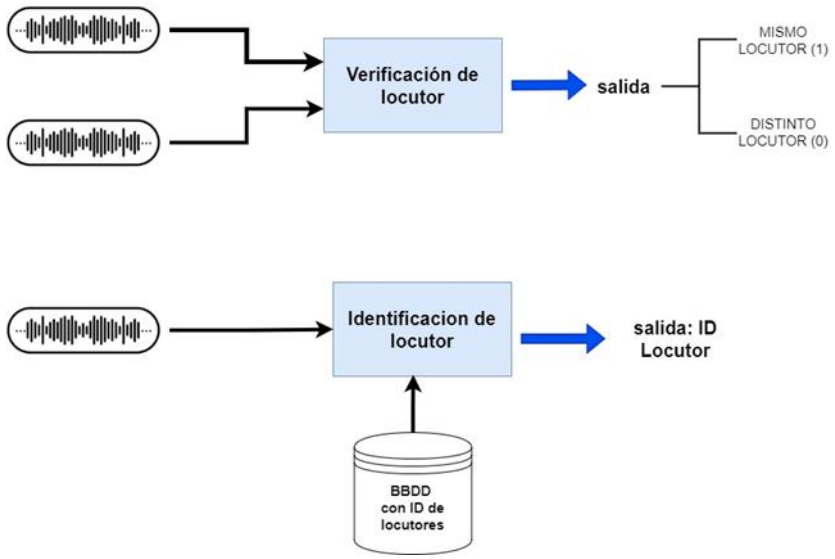

Figura 2-1: Esquema de verificación e identificación

La tarea de identificación de locutor se basa en reconocer a un sujeto a través de una muestra voz dentro de un conjunto de locutores previamente identificados. A su vez, la tarea de verificación de locutor consiste en comprobar si dos muestras de audio: una con un locutor identificado y otra con un locutor desconocido, pertenecen o no al mismo sujeto.

El primer paso que hay que realizar para poder llevar a cabo estas tareas, consiste en la extracción de las características. Esta extracción de características consiste en transformar el archivo *wav* en una secuencia de vectores que representan, las características acústicas a corto plazo, con estos vectores son con los que trabajamos **¡Error! No se encuentra el origen de la referencia.**. Los algoritmos más empelamos para realizar la extracción de características de la voz son: LPCC (*Linear Predictive Cepstral Coefficients*), MFCC (*Mel Frequency Cepstral Coefficients*) y MFB (*Mel-filter bank outputs*). MFCC son coeficientes calculados en una frecuencia basada en la percepción auditiva, por su parte, tanto los LPCC como los filtros Mel de los MFCC, son coeficientes basados en una predicción lineal que se asemeja al sistema auditivo humano [4].

En los sistemas actuales basados en redes neuronales, a las muestras de audio, se les aplica el banco de filtros en escala Mel (MFB), y, la salida de estos filtros (las características extraídas), son los datos de entrada de la red.

A partir de estos vectores de características podemos emplear diferentes técnicas para las tareas de verificación e identificación de locutores, como por ejemplo los modelos de aprendizaje automático (*maching learning*) y su extensión de aprendizaje profundo (*deep learning*).

Dentro de los modelos de aprendizaje automático podemos destacar los modelos de mezcla de gaussianas (GMM y GMM-UBM) y Supervectores. Los modelos GMM entrenados con los vectores de características, generando un modelo especifico por locutor; por el otro lado los modelos GMM-UBM está formado por con *Universal Backgorund Model* (UBM) el cual resulta ser un modelo general de todos los locutores (representando la variabilidad de la voz), y, el modelo GMM de cada locutor.

Los supervectores surgen a raíz de los problemas que se presentan en los modelos GMM y GMM-UBM. Un supervector resulta ser un vector de alta dimensionalidad que contiene la media de las gaussianas de un modelo de locutor [5][6][7].

Como los supervectores contienen información redundante para la tarea de reconocimiento de locutor debido a su alta dimensionalidad y a su origen de un modelo común (UBM), usarlos requiere un alto coste computacional. Como respuesta a este problema, surgen los modelos de variabilidad total (*total variability*, TV [8]), y, con ellos los i-Vectors.

2.1 Modelos de variabilidad total (TV)

La variabilidad total (TV) selecciona las dimensiones que contienen una mayor variabilidad con respecto al conjunto de datos sobre el que se entrena [8]. Con la reducción del espacio vectorial, pasamos a tener los i-Vectors estos albergan la información sobre el canal o la sesión (entendemos esto como los datos que no son per se las características del hablante), y, las características inherentes a los locutores.

La técnica de PDLA (*Probabilistic Linear Discriminant Analysis*) se emplean en los i-vectors para compensar la variabilidad que se genera en la sesión o canal. PLDA realiza las siguientes tareas: diferenciar entre la información del locutor y la del canal, busca modelar la variabilidad que se produce entre locutores diferentes (interlocutor) y entre las distintas muestras del mismo locutor (intra-locutor), de esta forma PDLA tiene como objetivo destacar la información que permite la diferenciación del locutor, y, minimizar la información que se genera debido al conjunto de muestras tomadas de un mismo locutor [9][10].

2.1.1 Redes Neuronales Profundas

Las redes neuronales es una de las técnicas que forma parte de las diferentes técnicas de aprendizaje automático, denominado: aprendizaje profundo o *Deep Learning*.

La finalidad de las redes neuronales es la resolución de problemas complejos, emulando el aprendizaje y resolución de los problemas que haría un ser humano a partir de unos datos.

Las redes neuronales están formadas por tres tipos de capas: capa de entrada, capas ocultas (dependiendo del tipo de red), y capa de salida, y cada capa cuenta con un número de neuronas definido en su diseño. Estas neuronas lo que hacen es aplicar funciones matemáticas sobre los datos de entrada, desde funciones lineales simples hasta transformaciones no lineales complejas sobre los datos.

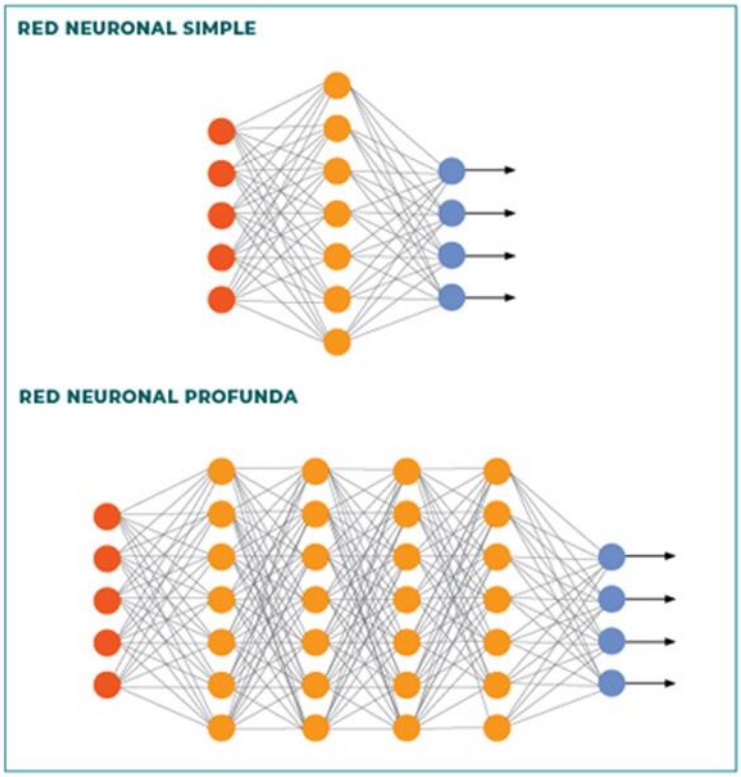

Figura 2-2: Esquema de una red neuronal [11]

Las redes neuronales profundas (DNN, *deep neural network*) se estructuran igual que las redes neuronales, con la diferencia, de que contienen muchas más capas ocultas, llevando a cabo transformaciones no lineales complejas sobre los datos. Este tipo de red se emplea para la resolución de problemas mucho más complejos como pueden ser sistemas de reconocimiento en los que hay que extraer características de los datos.

En la Figura 2-2 mostramos un ejemplo visual de la diferencia de las redes neuronales simples y redes neuronales profundas. La capa roja, en ambas redes se trata de la capa de entrada; la capa azul representa la capade salida que contiene tantas neuronas como clases tengamos en la red; y, por último, las capas naranjas representan las capas ocultas de la red, en función de la complejidad que necesitemos, tendremos más o menos capas ocultas. El actual estado del arte que resuelve las tareas de identificación y verificación de locutor se basan en sistemas de extracción de emebddings y X-Vectors, mediante el uso de redes neuronales profundas.

2.1.2 DNN-embedding (X-Vector)

Los embeddings es transformación de los vectores de alta dimensionalidad a una más reducida, de manera que los datos más relevantes se mantengan en este vector reducido. Estos embeddings son realmente útiles debido a que los vectores resultantes que contienen información clave sobre la secuencia de entrada, además, si hay datos similares o iguales, estos vectores van a ser muy similares.

Estos vectores nos permiten que podamos tratar de manera más eficientes estos datos, y, son comúnmente usados en tareas como clasificación de emociones, reconocimiento de idioma, reconocimiento de locutor, clasificación de texto, etc.

Por su parte, los X-Vectores son embeddings extraídos de redes neuronales recurrentes, como la que emplearemos en este trabajo, o, redes que cuentan con varias capas de TDNN (*time delay neuronal network*). Este tipo de redes o capas, lo que hacen es modelar los atributos invariantes en el tiempo que encontramos en los embeddings, para posteriormente extraer la media y desviación de los datos temporales (extraída de las anteriores capas), condensando esta información temporal, obteniendo finalmente los X-Vectors de dimensión fija independientemente de la longitud de la secuencia de entrada.

En [12], se comparan el uso de los sistemas de redes neuronales que emplean los sistemas de X-Vectors frente a los sistemas i-Vectors cuando se emplea un aumento de los datos. Dentro del estudio, se demuestra que el uso de los sistemas de X-Vector para el reconocimiento de locutores cuando contamos con un aumento de los datos (*data augmentation*) para mejorar el rendimiento, es mucho más efectivo que el uso de sistemas i-Vectors, y, presentan una mayor efectividad en muestras de corta duración (aproximadamente de 3 segundos).

Por este motivo, dentro de los sistemas de reconocimiento de hablante, se emplean redes neuronales profundas X-Vectors, aunque hayan mejorado la precisión en este tipo de tareas, la tasa de error aumenta con estos sistemas, cuando entre las grabaciones las condiciones como el ruido, la frecuencia de muestreo, etc., no son coincidentes. En este tipo de casos, un estudio realizado en 2020 [13], emplean *Siamese X-Vector Reconstruction* o en sus siglas, SVR. Demuestran que para este tipo de situaciones donde la variabilidad del entorno de las muestras es cambiante, mejoran la tasa de precisión frente a los X-Vectors.

2.2 Neutralidad en sistemas de aprendizaje automático (Fairness)

Para aplicar el concepto de neutralidad dentro del aprendizaje automático, previamente tenemos que explicar este concepto. La neutralidad es la idea de no favorecer a ningún grupo, idea u opinión; a diferencia de la imparcialidad, donde solo se evalúan los resultados objetivo y no la subjetividad.

Aplicar este concepto dentro del aprendizaje automático implica que los sistemas no pueden favorecer a un grupo poblacional frente a otro, cuando sesgamos en función de una característica. Cuando nos referimos a favorecer, nos referimos que el sistema tiene una

mayor precisión con un grupo frente al otro, o, este grupo favorecido presenta una menor tasa de error bastante significativa.

Al emplear cada vez más las técnicas de aprendizaje automático para realizar tomas de decisiones, el problema de la posible falta de neutralidad por parte de estas técnicas ha creado una línea de investigación prolifera en los últimos años. En estas investigaciones, se proponen distintos enfoques y métodos para conseguir paliar este problema, la mayoría se enfoca en el preprocesamiento de los datos, o, en el post-procesamiento de los datos [14][15][16][17][18].

Destacamos la técnica del preprocesamiento de datos ya que es una de las que emplearemos en uno de los experimentos descritos en el CAPÍTULO 4. El objetivo propuesto de la investigación [15], es paliar la falta de neutralidad dentro del conjunto de datos de entrenamiento modificando la forma en la que distribuimos las características sensibles (acento, edad, género, nacionalidad, raza…) por las que sesgamos.

2.2.1 Neutralidad en sistemas de reconocimiento de locutor

Al igual que sucede con los sistemas que emplean imágenes de las caras de las personas para realizar la identificación de personas, en los sistemas de reconocimiento basados en voz, han aumentado su uso tanto a nivel comercial como en aplicaciones de ciencias forenses y seguridad.

También en los sistemas de reconocimiento basados en voz, se han planteado los mismos problemas que surgen y sigue surgiendo con los sistemas de reconocimiento facial, la falta de neutralidad que hay entre distintos grupos poblacionales. Este problema puede ser debido a la falta de presencia de distintos atributos sensibles o una presencia reducida en comparación con otros [19][20][21][22]. De los estudios leídos, destacamos dos de ellos: el artículo [19] y el artículo [21].

En la investigación [19] llevan a cabo un experimento donde se emplea la base de datos FairVoice (base de datos que en este trabajo también vamos a emplear y que explicaremos en el CAPÍTULO 3), los autores plantean una posible solución para el problema de la falta de neutralidad mediante el preprocesamiento de los datos.

Para ello llevan a cabo un estudio previo de los datos que hay en la base de datos, tomando la decisión de emplear las etiquetas de edad y género para sesgar a los locutores en distintos grupos poblacionales. Dentro de la base de datos de FairVoice, encontramos distintos conjuntos de locutores en función de su idioma, pero, solamente presentan suficientes locutores el grupo de habla inglés y habla hispana para realizar un entrenamiento balanceado. Los autores dividen a los grupos en función de estas características: *male*, *female*, *junior* y *senior*.

La idea principal es realizar un entrenamiento donde estas 4 características que puede generar problemas de sesgo se encuentren dentro de la base de datos empleada para generar el entrenamiento y la evaluación de forma balanceada.

Los autores determinan que el empelar técnicas de equilibrar los datos, no termina de resolver los problemas de falta de neutralidad en los modelos pre-entrenados. También plantean la idea de realizar un nuevo modelo entrenando con los datos balanceados de

ambos idiomas, aunque el resultado no consigue unas mejoras significativas, dando lugar a un mejor nivel de neutralidad en español entre los grupos frente al inglés.

Por otro lado, en el artículo [21] analizan y evalúan el emplear una red de fusión adaptada a los grupos demográficos (GFN) – en el artículo solo generan dos grupos, *male* y *female* – centrándose en la fusión de los scores para abordar el problema de falta de equilibrio en las bases de datos. Para ello generan tres tipos de modelos: un modelo general, un modelo entrenado únicamente con locutores masculinos y otro modelo entrenado con locutores femeninos; se centran en general estos 3 modelos para extraer las características claves dentro de cada grupo demográfico.

Los resultados mostrados en el artículo [21], donde se estudió el efecto de la equidad de precisión sobre un grupo desequilibrado [22], muestran que el modelo propuesto logra reducir las faltas de precisión entre los grupos. Las diferencias de tasas de error igual entre los grupos sesgados y modelo general eran inferiores a 0.1, lo que indica una mayor neutralidad en el modelo. Este valor lo han obtenido a partir de la media de la disparidad entre las tasas de error (*Disparity Score*, DS), al igual que en otros experimentos que han empleado la misma técnica, por este motivo vamos a tener en cuenta esta métrica, explicada en el CAPÍTULO 3.

2.2.2 Calibración

En los estudios [2][23], se lleva a cabo una investigación sobre una de las posibilidades de las faltas de neutralidad de los sistemas, la falta de calibración en los datos que se empelan en para sistemas de reconocimiento de locutor, identificación de idioma, etc.

En ellos, se analizan el rendimiento de los sistemas con un conjunto de distintos acentos en inglés, demostrando que la precisión del sistema varía en función de este sesgo con problemas de calibración en los acentos que se encuentran de manera minoritaria en el sistema. Pero, cuando realizan uno de los experimentos entrenando de manera equitativa, se produce una mejora de la calibración y la neutralidad entre los grupos demográficos.

Para evaluar la calibración se empleó la métrica: *Calibration Loss* o pérdida de calibración (Cllr). Esta métrica no indica directamente si el sistema que estamos evaluando esta calibrado o no, pero indica cuánta perdida de la calibración se está produciendo en el sistema.

Con esta métrica se evalúan la calidad de la calibración: cuanto más cercanos al cero sean estos valores, mejor calibración hay en el sistema. Por el contrario, un Cllr alto puede indicarnos que el sistema estos experimentos problemas de calibración, lo que puede afectar directamente como indican con los resultados obtenidos, en la equidad del sistema.

3 Entorno experimental

En este capítulo explicaremos el entorno en el que hemos realizado las pruebas, explicaremos y detallaremos las bases de datos que hemos utilizado para el entrenamiento, los repositorios o herramientas para la generación de los modelos, y, que métricas usamos para evaluar estas pruebas.

3.1 Métricas de evaluación

En este apartado, vamos a explicar las 5 métricas que hemos utilizado para evaluar nuestros sistemas. Estas métricas son: *Equal Error Rate* (EER), *Minimum of the Detection Cost Function computation* (MinDCF), *Claibration Loss* (Cllr) y *Disparity Score* (DS).

Como se explica en el trabajo [2], la tasa de error igual (EER, explicada en apartado 3.1.1) es una medida poco eficiente al evaluar la falta de neutralidad en estos sistemas porque muchas de las bases de datos empleadas están desbalanceadas, por eso incluimos las métricas Cllr y DCF para añadir más información.

3.1.1 *Equal Error Rate* (EER)

La métrica EER (en sus siglas en inglés) o tasa error igual, es un calor comúnmente utilizado en sistemas biométricos. Esta variable indica el punto donde la tasa de falsos positivos (FP) y los falsos negativos (FN) es la misma.

En este trabajo, los falsos positivos se producen cuando el modelo clasifica como mismo locutor, los audios de dos locutores diferentes. Por otro lado, los falsos negativos se producen cuando el modelo clasifica como dos locutores distintos, los audios de la misma persona.

El EER o tasa de error igual será una de las métricas fundamentales para evaluar los sistemas, como se empleó en el trabajo previo del que partimos [1]. Las diferencias notorias de tasa de error igual que se producen en los grupos sesgados por una o varias características protegidas, pueden llevar a que el modelo generado presente una falta de neutralidad (*fairness*), favoreciendo uno o varios de los grupos que se están comparando.

Esta métrica la usamos para determinar la existencia de patrones dentro de los sesgos que producen faltas de neutralidad en los sistemas, igual que en los trabajos de guía y ya hemos mencionado [1][19][20][21][22].

3.1.2 *Calibration Loss* (Cllr)

La métrica *Calibration Loss* (Cllr) o la pérdida de calibración, se emplea para evaluar la calibración de modelos estadísticos, que se emplean para conseguir maximizar la verosimilitud del modelo.

Esta función en el ámbito de verificación de locutor se empleaba de forma conjunta con los modelos GMM y GMM-UBM, para determinar si las muestras de audios de dos locutores estaban relacionadas (mismo locutor) o no. Actualmente, algunos estudios [24] emplean esta función para mejorar los parámetros de las redes neuronales empleadas en verificación de locutor (*Speaker Verification*, SV), midiendo no solo el poder discriminativo del sistema (dentro de esta clase de sistemas y modelos, aumentar el poder de discriminación implica conseguir una mayor tasa de acierto), sino también la calibración.

En el ámbito de este trabajo, empleamos Cllr para determinar de la misma manera que lo emplean en el estudio ya mencionado [2], y como detallamos en el CAPÍTULO 2. El Cllr no indica directamente si tenemos un sistema calibrado o no, pero si nos indica si la falta de calibración (valores de Cllr altos) está contribuyendo a la falta de equidad del sistema.

3.1.3 *Minimum Detection Cost Function computation* (MinDCF)

La función de coste o DCF se emplea para medir el rendimiento del modelo y cuantificar el coste de decisión, esta función sirve para evaluar el rendimiento de un sistema. La función de coste se calcula como se presenta en la Ecuación 3-1:

$$\boldsymbol{DCF = ProbT * FNR * CostFN + (1 - ProbT) * FPR * CostFP}$$

Ecuación 3-1: fórmula DCF

Donde los parámetros:

- ProbT: probabilidad de que un objetivo ocurra a priori
- FNR: tasa de falsos negativos
- FPR: tasa de falsos positivos
- CostFN: coste de un falso negativo
- CostFP: coste de un falso positivo

El valor mínimo de la función de coste o MinDCF, es el valor que se obtiene al ajustar el umbral de decisión optimo que minimiza este coste. De forma conjunta, el valor DCF y el valor MinDCF se pueden emplear como una medida de calibración y discriminación, indicando como de acertado ha sido la estimación de los parámetros (pesos y bias).

En el contexto de este trabajo, el valor mínimo de la función de costes o MinDCF se emplea como métrica para evaluar si hay sesgos o no dentro del modelo. Cuando comparamos el valor de dos grupos divididos por una característica, al comprar ambos valores, podemos presentar una hipótesis sobre una tendencia a favorecer a un grupo o no.

3.1.4 Disparity Scores (DS)

La disparidad de puntuación o DS (en sus siglas en inglés), es una medida que podemos observar en que se emplean en las investigaciones [19][21]. La Ecuación 3-2 representa como se calcula la disparidad de puntuación entre dos grupos sesgados en

función de una característica sensible. Emplearemos esta métrica para determinar cuanta diferencia se presenta en los distintos grupos poblacionales, y, determinar si el DS es significativo o no.

$$DS = |EER_{G1} - EER_{G2}|$$

Ecuación 3-2: fórmula *Disparity Score*

En ambos artículos [19][21], para marcar un límite de este valor y determinar la existencia de *fairness*, toman la media los grupos evaluados, un procedimiento que vamos a seguir en este trabajo también. Todos los valores de DS que superen la media se consideran que son valores donde sí hay una falta de imparcialidad.

3.2 Bases de datos

Las dos bases de datos que vamos a emplear en este trabajo son las mismas que utilizamos en el trabajo del que partimos [1]: la base de datos VoxCeleb [25] y la base de datos Mozilla Common Voice [26], de la cual, usaremos el subconjunto de FairVoice.

3.2.1 VoxCeleb

La base de datos VoxCeleb se emplea para evaluar y comprar los sistemas del estado del arte actual. VoxCeleb es un conjunto de datos audiovisuales formado por clips cortos extraídos de entrevistas a personas famosas que se encuentran en YouTube. Esta base de datos cuanta con dos subconjuntos:

- VoxCeleb1: que tiene una cantidad total de 1251 locutores, y, más de 150.000 muestras de audios. Este subconjunto se emplea para la evaluación del sistema generado.

- VoxCeleb2: que tiene una cantidad de 6112 locutores, y presenta más de 1.000.000 de muestras en total. Este subconjunto es que se emplea para generar el entrenamiento de estos sistemas.

Ambas bases de datos presentan un archivo con la información de cada locutor, mientras que el archivo con los metadatos de Voxceleb2 solo presenta el identificador de cada locutor y el género al que pertenecen, el archivo de metadatos de Voxceleb1 nos da la información de género y nacionalidad. Como toda la base de datos resultan ser audios extraídos de entrevistas realizada a los locutores presentes en inglés, tomamos el dato de nacionalidad para determinar los distintos acentos que se generan al hablar inglés.

En las Figuras 3-1, 3-2, 3-3 y 3-4, se ha representado la distribución de los locutores sesgando por estas características. Con la información obtenida, queremos indicar: primero, que la base de datos no se encuentra balanceada y, por ende, podría ser

una de las causas de la falta de neutralidad, y segundo, mostrar los grupos poblacionales de los que generaremos subconjuntos de locutores distintas para evaluar.

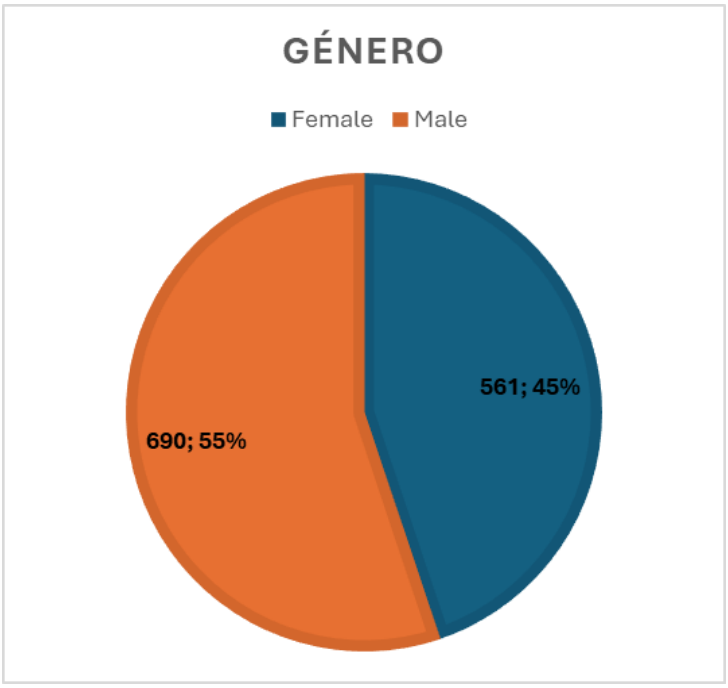

Figura 3-1: Distribución género VoxCeleb

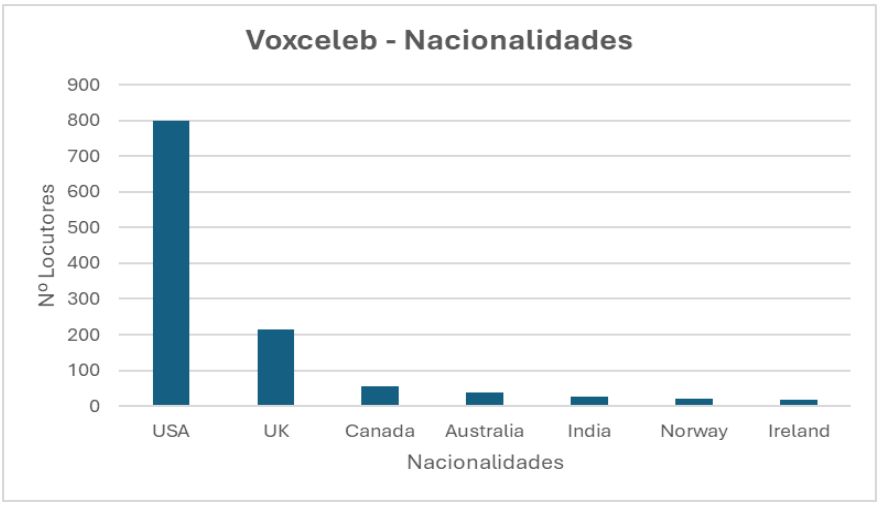

Figura 3-2: Distribución nacionalidades VoxCeleb - Pt1

Figura 3-3: Distribución nacionalidades VoxCeleb - Pt2

Figura 3-4: Distribución nacionalidades VoxCeleb - Pt3

3.2.2 Mozilla Common Voice: FairVoice

Como ya hemos mencionado al inicio, emplearemos el subconjunto FairVoice, que se divide en distintos idiomas. Como se encuentra explicado en los trabajos previos de los que partimos [1][20], utilizaremos solo los datos de locutores angloparlantes e hispanohablantes, ya que contienen un número adecuado de locutores para poder realizar un conjunto balanceado de los datos (esta acción la detallaremos en el CAPÍTULO 4, en el apartado 4.3).

En esta base de datos a diferencia de VoxCeleb, se detallan mucho más a fondo el tipo de acento que se presenta en un idioma, y también una característica más: la edad de los locutores. La edad de cada locutor se encuentra parametrizada en función a la década a la que pertenecen, por ejemplo, si un locutor tiene 45 años, se le parametrizará con la etiqueta de "*fourties*". En las siguientes figuras (Figura 3-5 hasta Figura 3-8), se representa la distribución de estos grupos poblacionales en la base de datos.

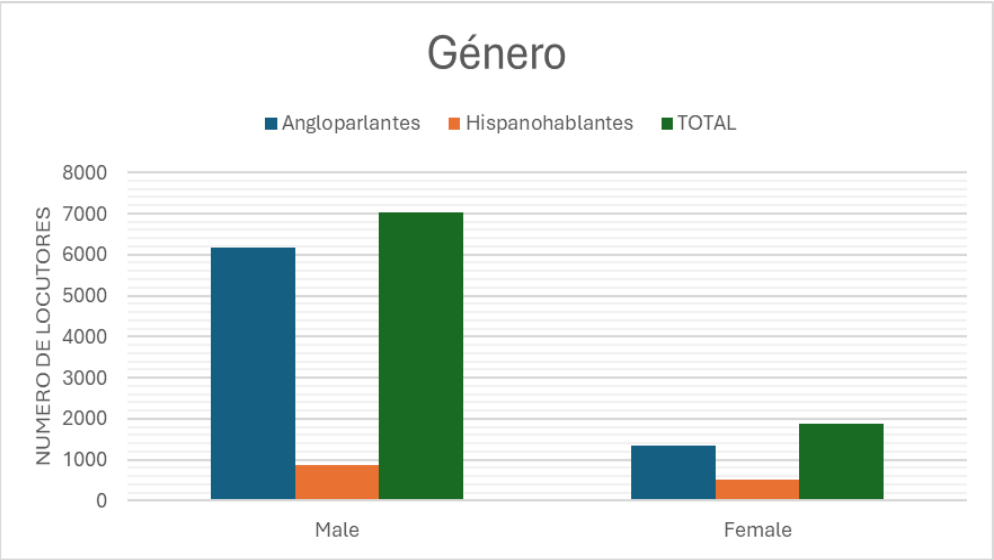

Figura 3-5: Distribución Género FairVoice

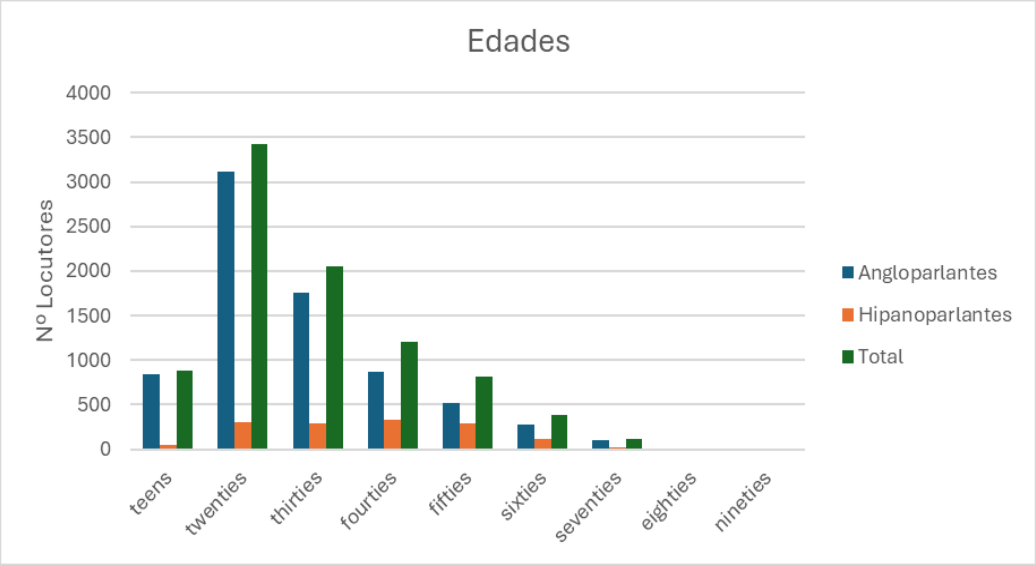

Figura 3-6: Distribución Edades FairVoice

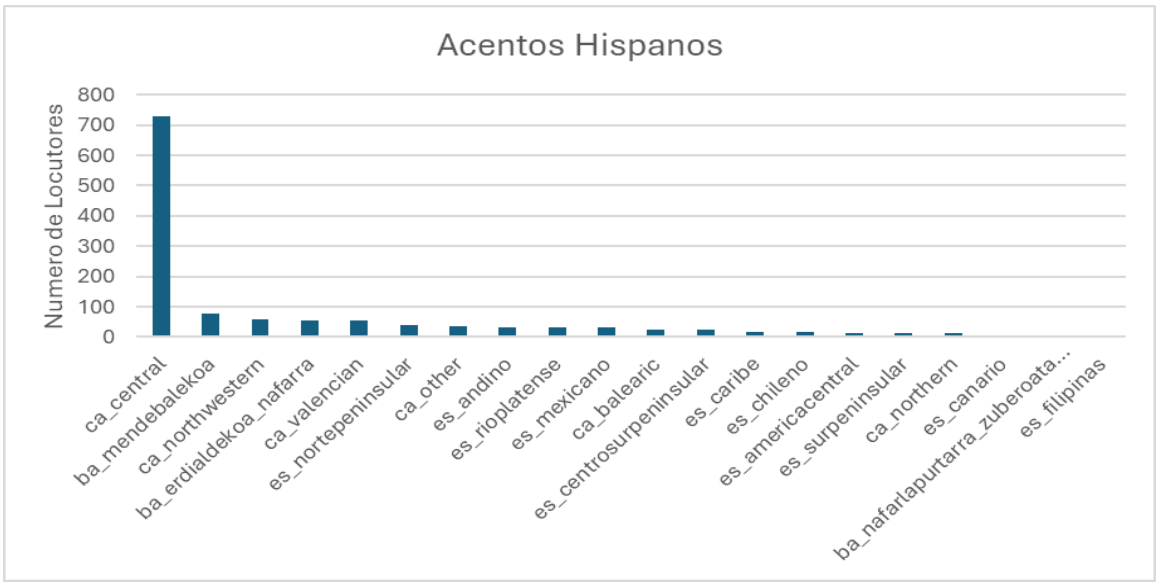

Figura 3-7: Distribución Acentos Hispanohablantes FairVoice

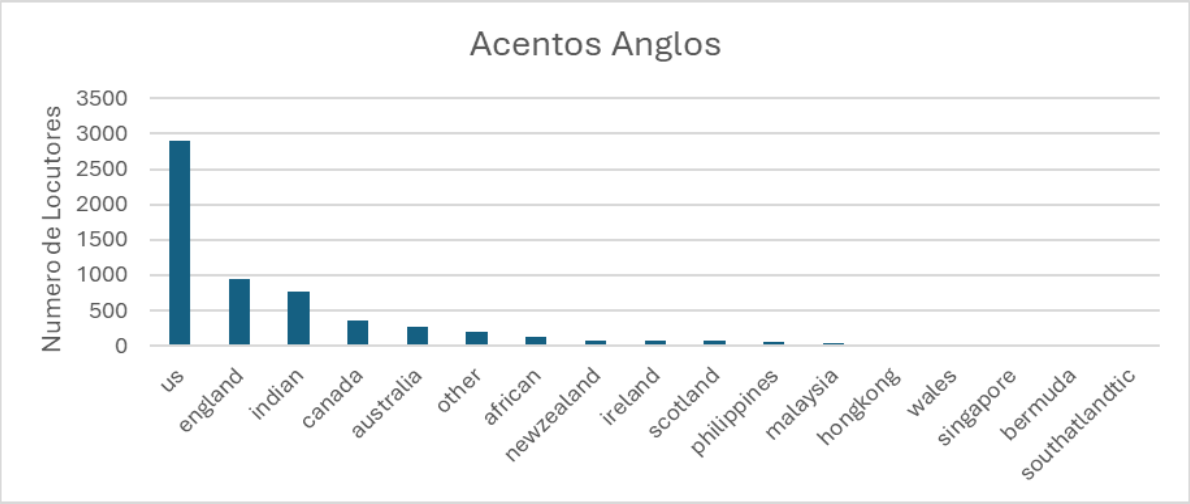

Figura 3-8: Distribución Acentos Anglohablantes FairVoice

3.3 Repositorios Software

Para llevar a cabo los experimentos vamos a emplear tres repositorios diferentes en este trabajo: el repositorio empleado en VoxCeleb [27], el repositorio presentando en el trabajo de FairVoice [20][28], y, el *toolkit* generado en el trabajo de calibración [2][29].

En el momento en el que estamos realizando este trabajo, VoxCeleb está considerando el *framework* para los sistemas del estado del arte. Cabe destacar que en este repositorio [27], no solo contamos con los códigos para realizar modelos de red neuronal profunda que emplea VoxCeleb, sino también los modelos baseline entrenados con la base de datos que tiene el mismo nombre [25]. Estos modelos baseline los emplearemos en uno de los experimentos que desarrollaremos en el CAPÍTULO 4.

Del repositorio presente en el trabajo de FairVoice [20][28], vamos a emplear el código que tienen para generar listas de entrenamiento y evaluación balanceados. Estas listas se generar a partir de la base de datos Mozilla Common Voice: FairVoice [26].

Por último, emplearemos el toolkit generado en el estudio de la neutralidad en los sistemas de verificación de locutor [2][29], como método para comprobar si la falta presente en los resultados de precisión de los grupos sesgados es debido a la presencia de sesgos reales, o, es producto de la falta de balance en las bases de datos.

Dentro de este conjunto de herramientas cabe destacar que en todos los códigos estamos empleando la biblioteca de aprendizaje profundo PyTorch, que al igual que utilizar redes neuronales profundas que utilizan los X-Vectors. Siendo el *toolkit* utilizado por la mayoría de los grupos de investigación y para el desarrollo de sistemas en el estado del arte.

Las modificaciones realizadas sobre los códigos originales se explicarán y detallarán en los distintos experimentos y, además, estos nuevos códigos generados y/o modificaciones están en el siguiente repositorio [30].

4 Experimentos, desarrollo y resultados

En este capítulo vamos a explicar los experimentos que hemos llevado a cabo en este trabajo y los resultados que hemos obtenido. En cada apartado, comentaremos los motivos del experimento llevado a cabo y daremos una explicación de las modificaciones que realizamos sobre el código base [27][28][29] del que partimos; ilustraremos mediante tablas y gráficos los resultados obtenidos, y, explicaremos estos mismos.

Como ya explicamos en el CAPÍTULO 3, concretamente en el apartado 3.1 MÉTRICAS DE EVALUACIÓN, emplearemos el conjunto de esas métricas para evaluar los sistemas y distintos resultados obtenidos.

4.1 Experimento 1: Evaluación del estado del arte con VoxCeleb

En este primer experimento que realizamos, vamos a evaluar las posibles faltas de neutralidad en uno de los modelos el estado del arte de los sistemas de verificación de locutor [25][27]. La idea surge de realizar una aproximación a los experimentos llevados a cabo en el artículo [2], en el cual se propone realizar una división entre los grupos poblacionales en función de la su nacionalidad.

Como ya explicamos en el CAPÍTULO 3, apartado 3.2 BASES DE DATOS, la base de datos VoxCeleb contiene audios de distintos personajes públicos. Estos audios son extractos de entrevistas que han realizado y se han extraído de clips de YouTube. Destacamos que todos los audios utilizados están en inglés, por lo que cuando hacemos alusión a nacionalidades estamos tratando los distintos acentos que hay presentes en la base de datos.

Para generar los subconjuntos sesgadas vamos a clasificar a los locutores en base a su género, como la mayoría de los artículos y del trabajo previo [1] hicimos en su momento, y, en base a su nacionalidad (acento). Creamos un script, que se pude encontrar en el siguiente repositorio [30], en donde empleamos el archivo que contiene la información de cada locutor para generar los subgrupos, en función de las características ya mencionadas, y, que evaluaremos.

Una vez tenemos los grupos sesgados creados, al contar ya con un modelo pre-entreando (*baseline.model*) no es necesario que realicemos el entrenamiento del modelo, y, solo necesitamos realizar unas modificaciones para poder obtener unos datos necesarios para la calibración.

En el trabajo de L. Ferrer [2], hay dos tipos de maneras de evaluar la calibración de los datos, una mediante los scores a la que denomina calibración PostHoc, y otra mediante los embeddings. Durante todo este trabajo no hemos logrado realizar la calibración mediante los embeddings como habíamos propuesto, y, con la información obtenida no es suficiente para poder realizar esta calibración, y, no contamos con ejemplos de cómo hay

que obtener ciertos elementos para obtener los resultados. Por este motivo, durante este experimento y los demás que presentamos, solo vamos a emplear la calibración basada en scores: denominada calibración PostHoc.

Con los datos de los *scores* y *labels* a nuestra disposición, tanto los obtenidos de la lista de prueba general, como los de las listas sesgadas, procedemos a realizar la calibración mediante estos scores, PostHoc. Para ello, empleamos los scripts obtenidos del trabajo [2] y que se encuentran en el repositorio [29].

Creamos una pequeña función para generar el archivo *npz*, que se nos solicita, en esta función empleamos los *scores* y *labels* guardados de la evaluación del sistema, los dividimos en dos grupos en función de la etiqueta correspondiente: positivo (puntuaciones correspondientes a verdaderos positivos) y negativos (puntuaciones correspondientes a verdaderos negativos), estos dos grupos se guardan en el archivo *npz* que se emplea para el entrenamiento de un modelo de calibración y la evaluación de los grupos.

Generamos así un pequeño script donde incluimos esta función, y las correspondientes de evaluación mediante scores y entrenamiento mediante los scores. Este script está adaptado a que recorra la carpeta donde tenemos estos datos y vaya generando las calibraciones de forma consecutiva, automatizando el proceso.

En la Tabla 4-1, plasmamos los valores de EER y MinDCF que hemos obtenido con la evaluación de las listas. El valor EER lo vamos a emplear para obtener *el disparity score* o DS de los grupos. La media de todos los valores DS vamos a emplearla como límite, como bien ya explicamos, este límite será una forma de considerar las faltas de neutralidad presentes en la base de datos para los distintos grupos poblacionales.

Todos los valores que superen ese límite, vamos a considerar que presentan una falta de neutralidad y en los que nos centraremos a la hora de estudiarlos. Mediante el MinDCF comprobaremos en los distintos grupos que presenten esta falta de imparcialidad, si es dado el este problema sucede debido a que el sistema no es capaz de discriminar correctamente a los locutores.

	General	Male	Female	USA	UK	Canadá	Spain	Norway	Mexico	Ireland	France
EER	2,179	2,502	1,47	2,172	2,249	5,797	3,516	1,495	2,757	1,42	1,205
MinDCF	0,169	0,181	0,133	0,174	0,152	0,14	0,109	0,094	0,184	0,158	0,102

Tabla 4-1: EER y MinDCF VoxCeleb

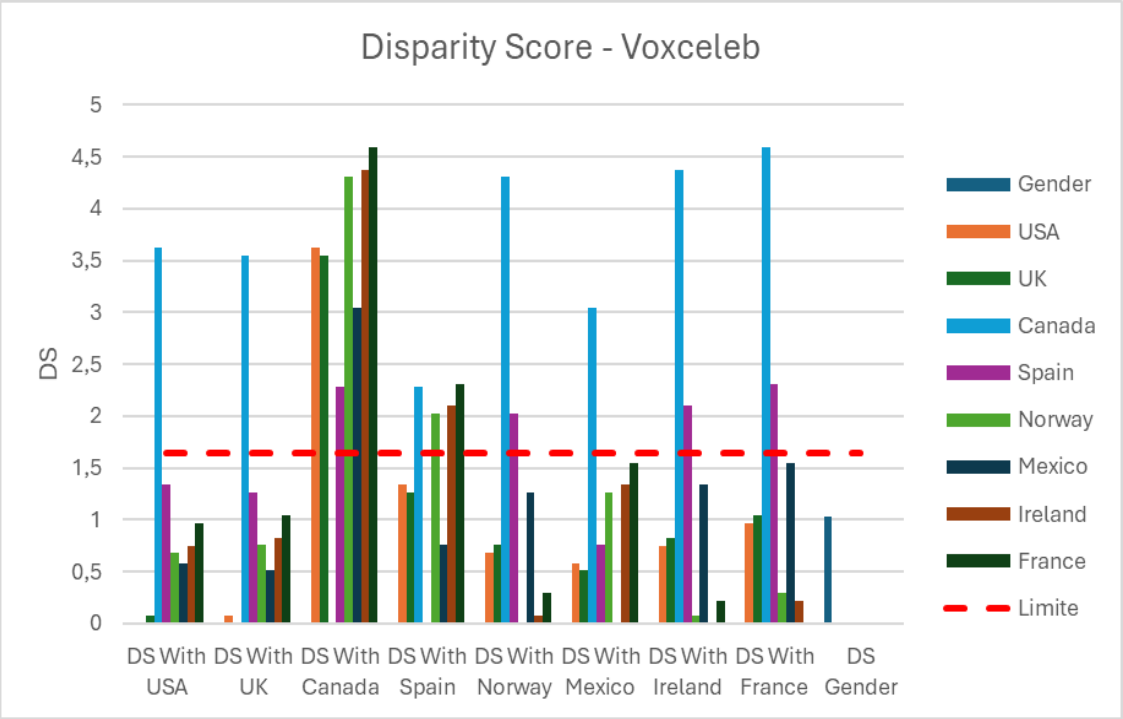

Figura 4-1: *Disparity Score* VoxCeleb

En la gráfica de la Figura 4-1 estamos mostrando los resultados del *disparity score* o DS, para generarlo, lo que hacemos es aplicar la Ecuación 3-1 con los distintos grupos. La fórmula se aplica entre dos grupos: para hacer el DS *Gender* empleamos solamente le grupo *male* y el grupo *female*; pero, para hacerlo los distintos DS de acentos, lo que hacemos es aplicar de fórmula con todos los grupos comparándolos con el grupo que el DS marca.

En el eje de las X tenemos los distintos grupos sobre los que queremos comparar al resto de grupos, y en el eje de las Y está representando el valor de la disparidad que existen entre ambos grupos.

Al estudiar con detenimiento los datos podemos decir que el grupo *Canada* presenta con el resto de los acentos una falta de neutralidad, y, entre los demás grupos, destacamos la falta de neutralidad presentada al comprar el grupo *Spain* con los grupos *Norway*, *Ireland* y *France*.

Ambos grupos, *Canada* y *Spain*, presentan valores de MinDCF por debajo de la media general de todos los grupos si lo empleamos como una medida, y, si empleamos como medida limite estándar el valor MinDCF obtenido con la lista general, también se encuentran por debajo de este valor. En la gráfica pintada en la Figura 4-2, representamos los valores MinDCF y estos límites establecidos para mostrarlo de manera más visual y que resulte más fácil comprender lo que decimos

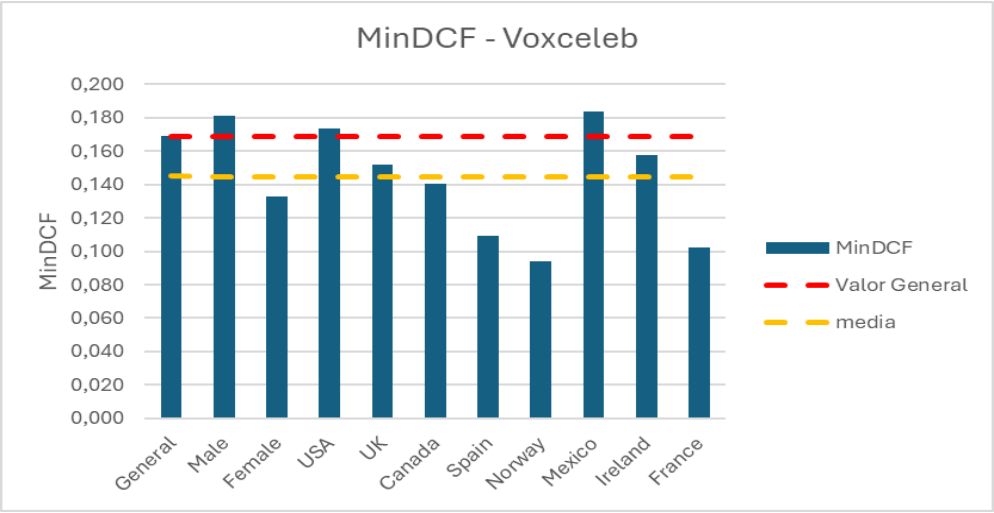

Figura 4-2: MinDCF VoxCeleb

Los valores de estos grupos (*Spain y Canada*) nos están indicando, que el modelo si está estimando bastante bien los pesos y bias para poder extraer las características específicas para estos grupos, y, lograr identificar a los locutores. Si determinamos que el problema no está viniendo del sistema, podemos plantear la idea de que el problema puede ser un problema de la calibración de los datos.

Para llevar a cabo la calibración mediante scores, preparamos el script que habíamos creado con anterioridad. De todos los datos que obtenemos de este script vamos a centramos en dos de ellos: el Cllr después de la calibración para ptar=0.1 y ptar=0.5.

El ptar (*Probability of Target Absent Rate*) es una medida empleada en los sistemas de reconocimiento de locutor que indica la probabilidad de que un locutor se encuentre fuera de la evaluación, en el contexto de este trabajo, se emplea como una medida de evaluación de las condiciones del sistema: un ptar bajo expone unas condiciones de presentar más locutores conocidos por el sistema y un ptar más alto indica que es más probable que los locutores no se encuentren dentro de la evaluación. Durante todos los experimentos vamos a emplear los mismos valores de ptar para llevar a cabo la evaluación de las pruebas.

En la gráfica que se observa en la Figura 4-3, se representan los valores de Cllr después de haber realizado la calibración del sistema con estos dos valores diferentes de ptar. Hemos incluido la media de los Cllr obtenidos por los diferentes grupos para poder tener una referencia visual, que ayude a la interpretación de los datos.

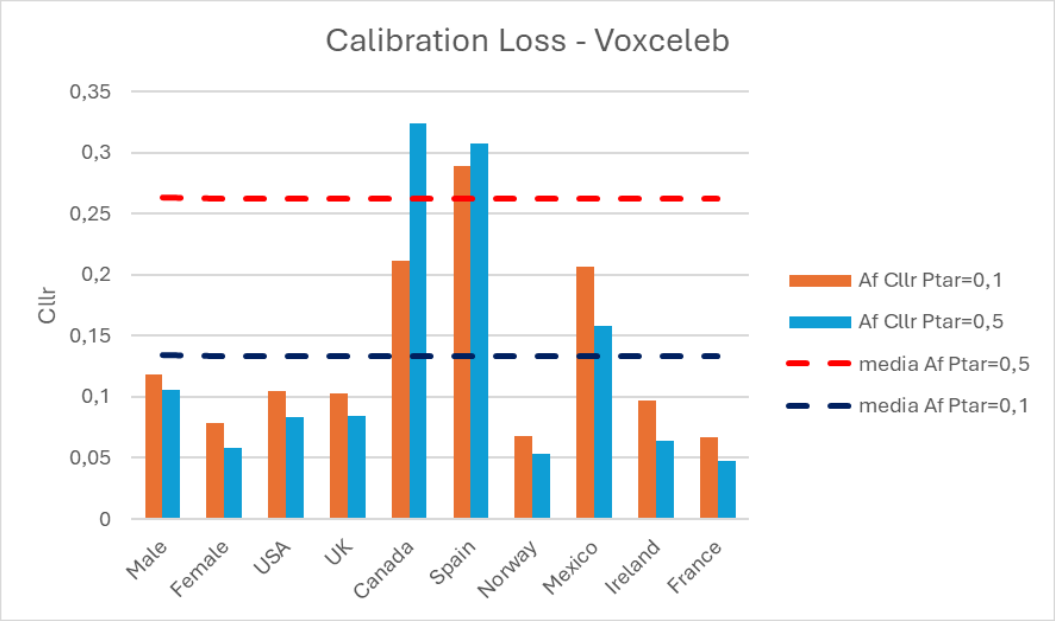

Figura 4-3: Cllr después de la calibración

En la Figura 4-3, podemos comprobar cómo tanto el grupo *Canada* como el grupo *Spain*, superan ambas medias, por lo que podemos intuir, que sus faltas de calibración en los datos están afectando a la falta de neutralidad del sistema.

Por otro lado, el grupo *Mexico* presenta un valor de Cllr (ptar=0.1) por encima de la media de ese grupo, pero el otro de Cllr (ptar=0.5) no, pudiendo ser interpretado que cuando la probabilidad de estar presente en las evaluaciones es mayor la calibración de sus datos no es tan buena como cuando la probabilidad de estar presente es menor.

Tras los resultados obtenidos y estudiados detenidamente, podemos teorizar sobre la falta de equidad que se presenta en los dos grupos principales (*Canada y Spain*), provienen de la falta de calibración de la base de datos, como se observa en la Figura 4-3, en lugar de suponer que esta desigualdad de precisión se produce por el sistema, como bien refleja la Figura 4-2.

Con estos resultados, y esa primera hipótesis, queremos comprobar como funcionaría el sistema entrenando con la base de datos de FairVoice, la cual, como ya se indica en el apartado 3.2 BASES DE DATOS, contiene más información de los locutores, con una base desbalanceada, y, realizaríamos un tercer experimento para comprobar la efectividad de emplear una base de datos balanceada.

4.2 Experimento 2: Generación y evaluación de un modelo no balanceado con FairVoice

En este segundo experimento, buscamos hacer una réplica similar del primer experimento llevado a cabo, pero empleando la base de datos FairVoice de forma no balanceada como se emplea en el trabajo [19], es decir, que, dentro de los grupos de edad y género, no haya presentes la misma cantidad de locutores.

Para replicar el primer experimento, emplearemos la misma arquitectura de red, y el mismo código que hemos empleado, con las modificaciones añadidas para la obtención de los datos para su posterior calibración.

Como ya se explicó en el apartado 3.2 BASES DE DATOS, en esta base de datos, los subconjuntos de español e inglés son los únicos que contienen suficientes locutores para poder llevar a cabo un entrenamiento balanceado, por lo que serán estos dos mismos subconjuntos los que vamos a emplear en este experimento.

Vamos a generar el modelo, empleando de forma conjunta ambos idiomas, esta decisión está fundamentada en la idea de conseguir no solo un modelo más robusto frente al ruido (al contar con más datos), sino también un modelo algo más universalizado, en lugar de dos modelos específicos para cada idioma.

Para generar los conjuntos tantos de entrenamiento como de evaluación, vamos a emplear el archivo metadata.csv que contiene la información de los usuarios: el identificador de cada uno de ellos, el número de muestras que hay presentes en la base de datos, el idioma, la edad, el acento y el género.

Con esta información, primero vamos a cribar a los locutores en función del idioma escogiendo únicamente los que hablen en español e inglés, y, después los cribamos en función del número de muestras que tenga cada locutor, escogiendo únicamente aquellos que tengan al menos 5 muestras de audios.

Los primeros dos conjuntos que generamos son las de entrenamiento y evaluación, esta última sin realizar ningún tipo de sesgado en función de las siguientes características protegidas: género, edad, acento e idioma. Para generar los subgrupos de género y edad, lo único que hacemos es dividir a los locutores en función de los respectivos valores que presenten.

Para dividir a los locutores por edad, en lugar de generar grupos sesgados en función de cada década de edad presente, vamos a dividirlos en dos grupos: senior y junior. Esta decisión viene motivada tras el análisis de la presencia tan desigualada de los grupos en la base de datos, como bien refleja el gráfico de la Figura 3-4.

Aunque completar este estudio de las desigualdades de precisión entre las edades, proporcionaría una información realmente útil para futuros trabajos, queremos tener un mínimo de locutores presentes en las evaluaciones para poder tener datos significativos, y esto no se podría lograr con ciertos grupos de edad. Esta causa de peso hace que tomemos la decisión de tomar el conjunto de las edades presentes, y convertirlo en un dato binario con los dos términos que hemos mencionado: *senior y junior*.

Dividimos el conjunto de edades siguiendo la misma lógica que presentan en el trabajo [19], emplear la etiqueta "*fourties*" como límite, todas las etiquetas de edad que se corresponda a una edad por debajo de 40 van a formar el conjunto junior, y, todas las etiquetas de edad que se corresponden a una edad igual o superior a 40 van a formar el conjunto senior. En la Figura 4-4, queda reflejado como quedaría la distribución de la edad con esta forma en el conjunto de la base de datos empleada para este experimento.

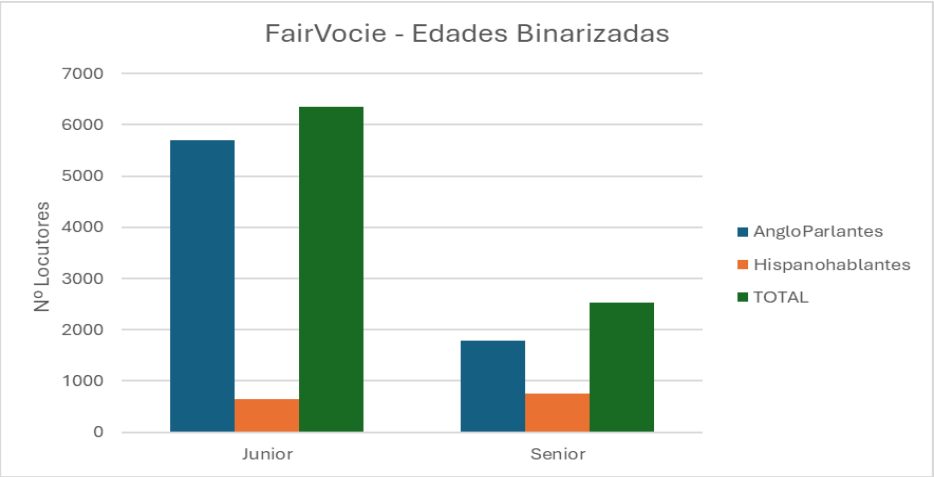

Figura 4-4: Distribución edades binarias – FairVoice

Para la división por idioma, lo que buscamos es replicar esa parte de los experimentos que llevamos a cabo en el trabajo [1] y que también se lleva a cabo en el experimento [19]. Queremos comparar cómo se comportan las características de género y edad, en función del idioma, y, con ello obtener más información sobre los resultados obtenidos de forma general. Si nos volvemos a fijar en la Figura 4-4, encontramos que hay más locutores senior hispanohablantes, que locutores junior hispanohablantes, un motivo más que refuerza la idea de realizar esta comparación de idiomas.

Para la generación de acentos, de todos los subgrupos generadas, únicamente vamos a escoger aquellas que presentan un número significativo de locutores para la evaluación. Como bien se puede observar en las gráficas de las Figuras 3-5 y 3-6, no todos los acentos disponibles en la base de datos cuentan con suficientes locutores para obtener unos datos significativos, por ese motivo, solo escogemos los que más representación tienen.

Con los subconjuntos generadas, y el modelo ya entrenado, procedemos a realizar las distintas pruebas de evaluación. Al igual que en el anterior experimento, emplearemos el EER de cada grupo para obtener el valor de *disparity score* de cada grupo comparado y el MinDCF para evaluar el sistema. Con las modificaciones del código original, obtenemos en cada evaluación los *scores* y *labels* que posteriormente se emplearán para realizar la calibración.

	General	Male	Female	Junior	Senior	Male Junior	Male Senior	Female Junior	Female Senior
EER	4,548	4,249	5,130	4,237	4,638	4,003	3,900	4,670	5,650
MinDCF	0,221	0,179	0,262	0,203	0,235	0,180	0,192	0,260	0,266

Tabla 4-2: EER y MinDCF – FairVoice

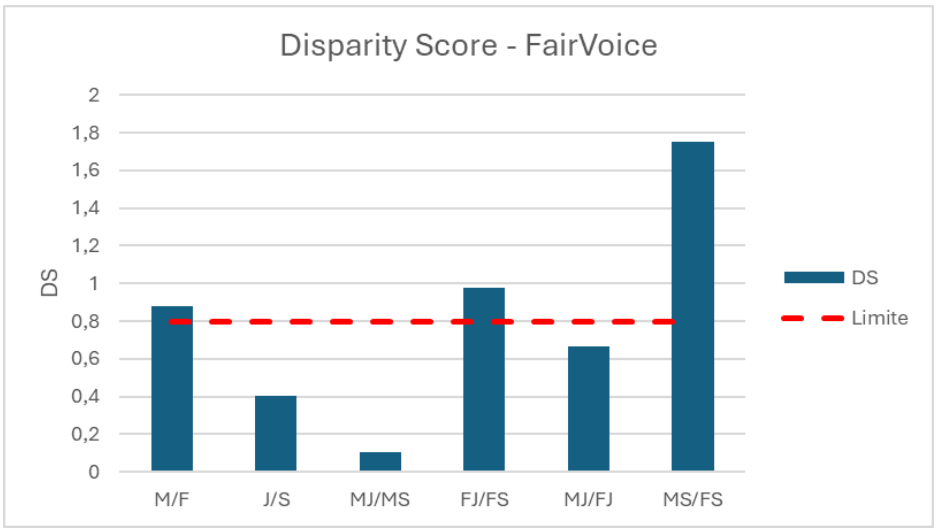

Figura 4-5: *Disparity Score* – FairVoice

En la gráfica de la Figura 4-5 estamos representando los valores del *disparity score* que hemos obtenido para esta primera prueba de evaluación, en donde dividimos a los grupos en función del género, la edad y una combinación de ambas. Como podemos observar en los resulatdos de la grafica, los valores del *disparity score* nos indican que entre los subgrupos sesgados por: género (male/female), los locutores femeninos por edad (female senior/female junior) y entre los locutores senior por género (male senior/ female senior) se estan produciendo sesgos.

Como realizamos en el primer experimento, entendemos que estas faltas de neutralidad pueden deberse: *a la falta de representación y calibración en los datos, o, a sesgos producidos por la falta de capacidad de discriminación del sistema*.

Por ello, primero vamos a realizar el cálculo de la perdida de calibración (Cllr) con los datos centrándonos en los grupos: *male, female, male senior, female senior y female junior*; debido a que encontramos que generan las diferencias de precisión notables.

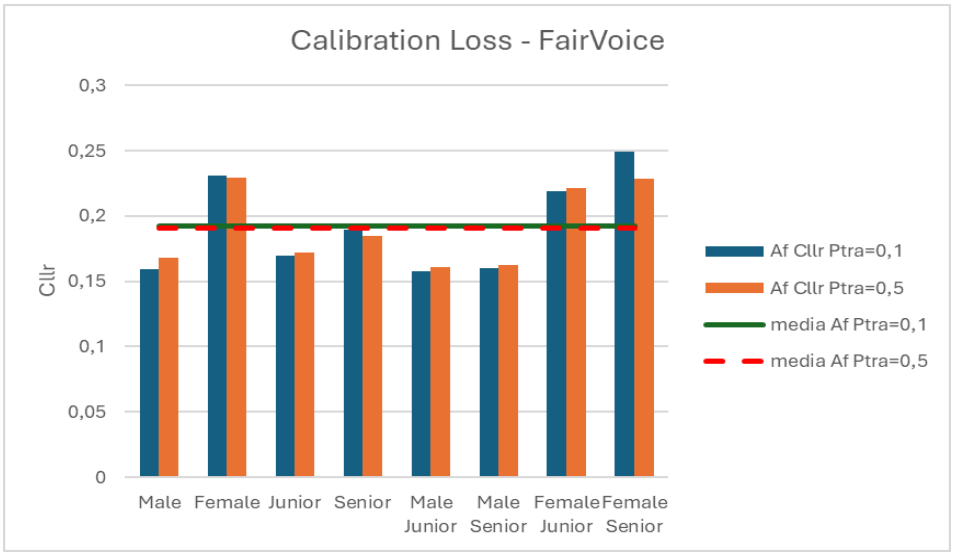

Figura 4-6: *Calibration Loss* – FairVoice

Para ambos resultados de la pérdida de calibración (el cálculo realizado con ptar igual a 0.1 e igual a 0.5) los grupos de *female, female junior y female senior* superan las medias correspondientes de cada resultado, como se expone la Figura 4-6.

Estos datos que nos están indicando que los problemas que están surgiendo son a raíz de la falta de calibración en los datos, y, al revisar el valor del MinDCF de los tres grupos, apreciamos que superan tanto el valor medio, como el valor que se genera en la evaluación de todo el grupo de test, como se ve en la gráfica Figura 4-7

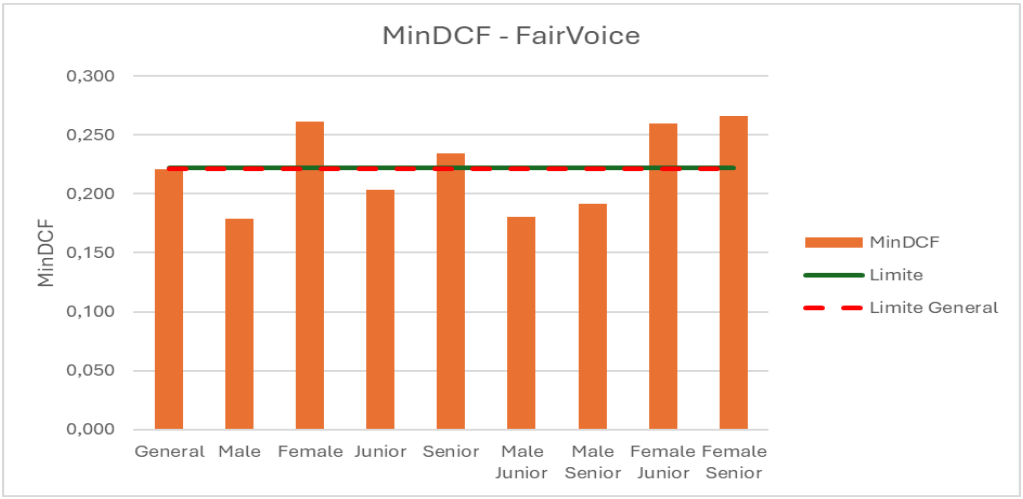

Figura 4-7: MinDCF subgrupos FairVoice

Nos planteamos la siguiente hipótesis: la baja representación en la base de datos de algunas características privadas (como es el caso de las etiquetas *female y senior* en esta base de datos) genera que el sistema tenga más dificultades para poder realizar una discriminación de los locutores que presenten esas características, y, con ello genere un sesgo.

	African	Australia	Canada	England	Ireland	Newzeland	Indian	US
EER	2,609	0,870	1,739	2,609	1,317	1,739	1,7391	2,609
MinDCF	0,104	0,017	0,200	0,070	0,096	0,113	0,02609	0,052

Tabla 4-3: EER y MinDCF – acentos ingleses

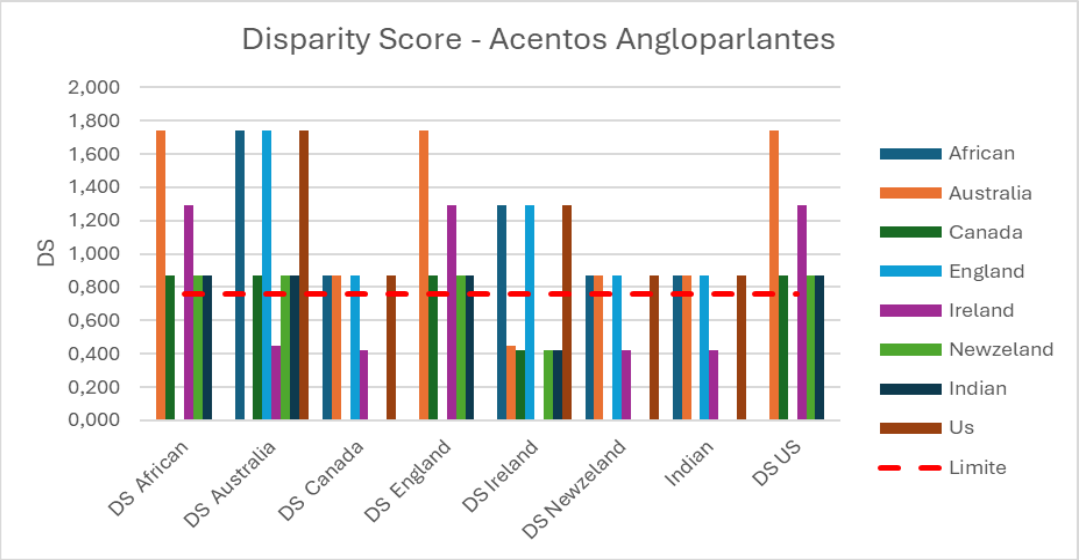

Figura 4-8: *Disparity Score* – acentos ingleses

Esta segunda parte del experimento, evaluamos los acentos del idioma inglés que presentan en los locutores de la base de datos. De los grupos que más destacan al superar la media son: *Australia* e *Ireland*; y, en menor medida, pero que se repiten en todos los casos: *Canada, England, African y US.*

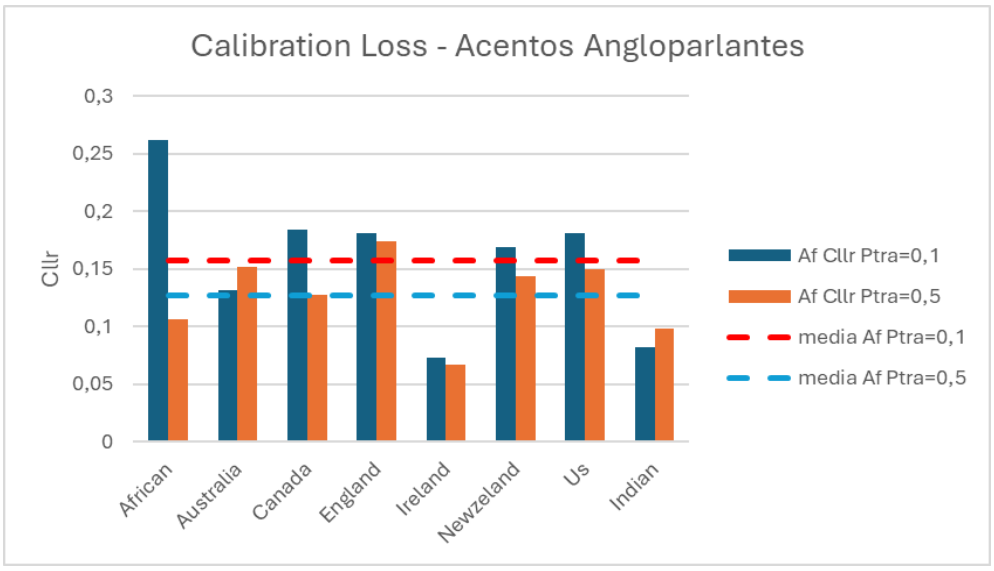

Figura 4-9: *Calibration Loss* – acentos ingleses

La pérdida por calibración nos está indicando que para los grupos: *African, England, US, Canda y Newseland*, la base de datos no está bien calibrada para esos grupos en uno o ambos casos. En los casos de *US y England* se tratan de los dos acentos de los que contamos con más locutores en la base de datos, por eso nos sorprende que, aunque su falta de calibración no es muy alta (como en otros casos), sí nos sorprende que estén por encima de la media. Con los datos obtenidos con esta prueba, no terminamos de esclarecer, si los problemas de falta de igualdad en la precisión se deben a los datos o es producido por el sistema. Vamos a realizar una comparación del MinDCF para intentar esclarecer este caso.

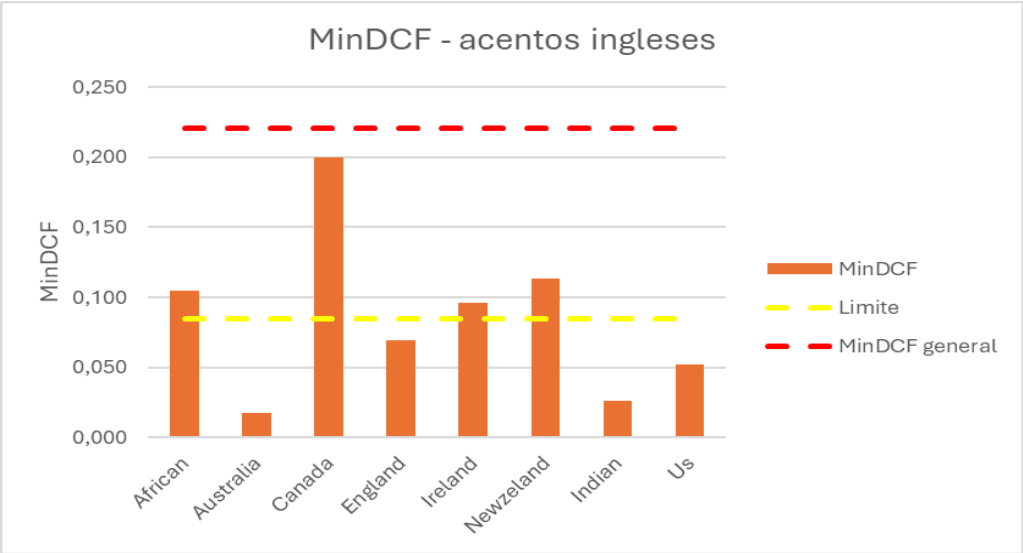

Figura 4-10: MinDCF - acentos ingleses

En la Figura 4-10 representamos los valores del MinDCF que obtenemos con los distintos acentos, hemos representado dos límites, uno con el MinDCF que obtenemos de la evaluación de todos los locutores, y otro con la media de este grupo en concreto. Comprando los datos, entendemos que al sistema le resulta más complicado discriminar a los locutores de los grupos *Canada*, *African y Newzeland*. Para estos tres casos, el problema podemos teorizar que proviene de su poca representación dentro de la base de datos.

	ba erdialdekoa nafarra	ba mendebalekoa	ca central	ca Nort Western	es rio platense	ca Other	es Norte Peninsular
EER	3,750	1,250	3,750	1,250	2,727	1,250	1,250
MinDCF	0,075	0,013	0,150	0,013	0,182	0,025	0,038

Tabla 4-4: EER y MinDCF – acentos españoles

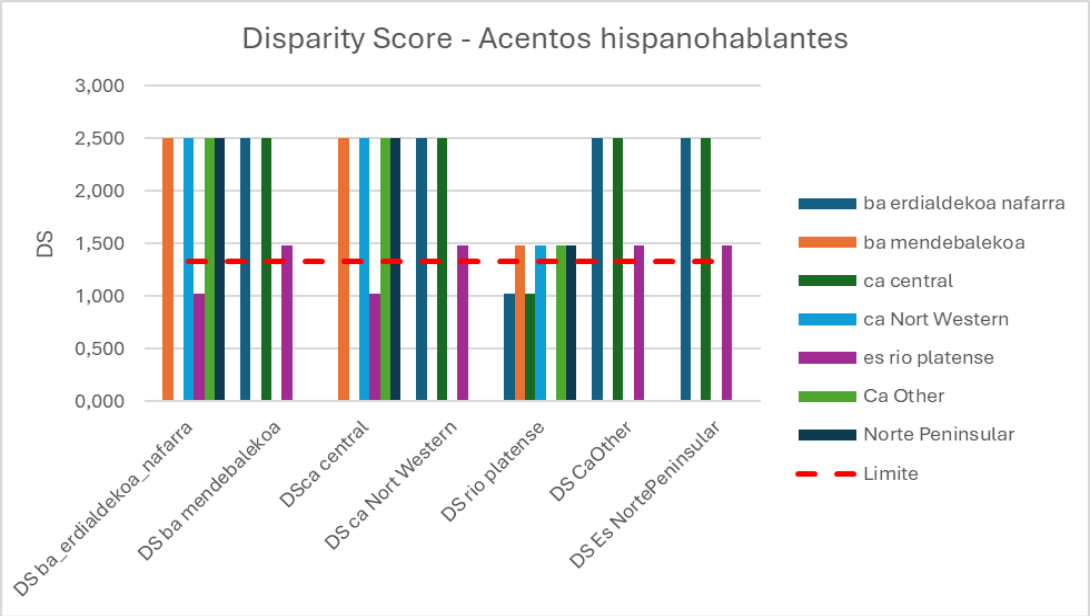

Figura 4-11: *Disparity Score* – acentos españoles

Este caso es similar a la anterior parte del experimento, estamos evaluando los resultados para listas con los locutores de habla hispana, comparando los distintos acentos que más se presentan en la base de datos. Si compramos los resultados de la Figura 4-11 y la Tabla 4-4, junto con los obtenidos tras la calibración, mostrados en la Figura 4-12, podemos determinar que, en este caso, las faltas de neutralidad vienen dadas por los grupos: *Ca_central*, *Ba_erdialdekoa_nafarra y es_norte_peninsular*.

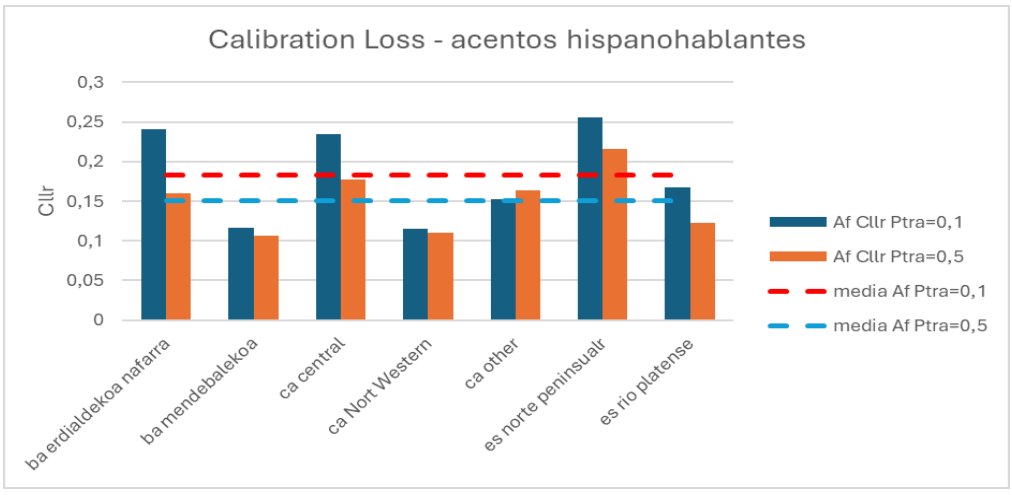

Figura 4-12: *Calibration Loss* – acentos españoles

Además, analizando los valores del MinDCF que obtenemos para este grupo, Figura 4-13 los acentos *Ca*_central y *Es_rioplatense* son los dos acentos que le resultan al modelo poder discriminar correctamente. Con la información obtenida, hemos generado dos teorías al respecto.

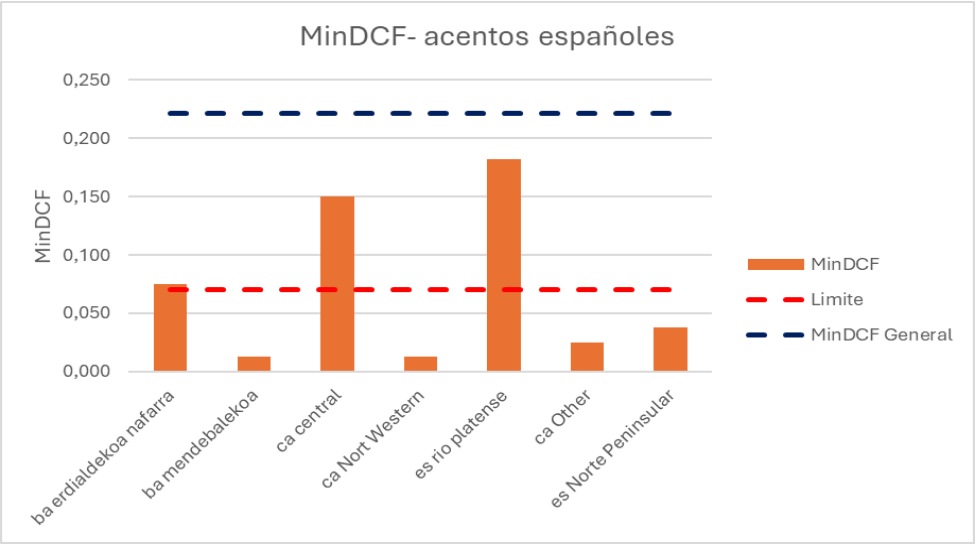

Figura 4-13: MinDCF – acentos españoles

La primera teoría, refuerza la ya propuesta en el EXPERIMENTO 1: EVALUACIÓN DEL ESTADO DEL ARTE CON VOXCELEB: la baja representación de algunas características en los datos del conjunto que empleemos genera problemas de calibración de estos mismos datos, que se ven reflejados en las faltas de equidad entre grupos sesgados en función de esa misma.

La segunda teoría se corresponde con los acentos, y los datos que tenemos. Partimos de la base que las características prosódicas y fonológicas (tono, acento, ritmo, entonación, timbre, intensidad…) forman parte de las características que extraemos para identificar o verificar a un locutor [31][32][33]. Estas características no solo varían entre idiomas, sino también entre los distintos acetos que existen dentro de un mismo idioma [34][35]. Partiendo de esa información y con los resultados obtenidos, generamos nuestra segunda hipótesis: *los locutores que tienen unos acentos más característicos de una región en concreto presentan ante el sistema una mayor facilidad para poder ser discriminados correctamente frente a los demás, y como resultado que esos locutores sean más fácilmente identificables*.

Un caso para ejemplificar esta segunda hipótesis son los acentos españoles: castellano central (*Ca_central*) y vasco (*ba mendebalekoa*), el primero se trata de un acento más generalizado en la península que abarcar distintas regiones de la meseta central, mientras que los locutores del segundo acento pertenecen a una región más reducida y el propio acento es mucho más identificativo.

Como resumen de esta segunda teoría, siempre y cuando el acento sea bastante más generalizado o neutro, puede llevar al sistema a tener más dificultades para poder discriminar a los locutores de ese grupo, aunque sea uno de los grupos con mayor representación de la base de datos. Consideramos la posibilidad de que los acentos, a diferencia de otras características privadas, no solo depende de su presencia en la base de datos sino también de su singularidad.

4.3 Experimento 3: Generación y evaluación de un modelo balanceado con FairVoice

En este experimento vamos a generar un modelo como el anterior, pero en lugar de emplear la base desbalanceada, vamos a utilizar el código proporcionado en la investigación [19][28], que genera conjuntos de entrenamiento y evaluación balanceados, es decir, contamos con la misma cantidad de locutores etiquetados como *male, female, senior, junior* y las combinaciones entre los grupos (*male senior, female senior, male junior* y *female junior*).

Al realizar este balance de locutores, pasamos de tener la cantidad que teníamos antes, a un número más reducido: 480 locutores por cada idioma, en total 960 locutores para entrenar este modelo. Destacamos esta información para establecer que tanto la tasa de error general, como la de los grupos sesgados por las características privadas va a aumentar, ya que contamos con menos variabilidad en la red. Esta reducción de locutores afecta a la cantidad de acentos que se van a presentar en la red y en los conjuntos de evaluación, por lo que contamos con un número reducido de estos a la hora de realizar las evaluaciones.

Como el *script* que genera las listas balanceadas no están adaptadas al formato de red de VoxCeleb, generamos unas funciones dentro de nuestros scripts de creación de las listas de FairVocie para adaptar las listas al formato que empleamos.

Cuando tenemos las listas de entrenamiento y evaluación adaptadas, procedemos a realizar las listas sesgadas en función del género, la edad y los acentos de cada idioma, el mismo sesgo que realizábamos en el 4.2 EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE. En lo relativo a la división de los locutores en función de la edad, vamos a seguir manteniendo el formato binario de *junior* y *senior*. Como esta explicado en el artículo [19] y en el anterior experimento.

En la Tabla 4-5 exponemos los valores de EER y MinDCF obtenidos con la primera evaluación de las listas de FairVoice, mientras que en la Figura 4-14 mostramos el resultado de realizar la formula del *disparity score* entre los grupos. En esta gráfica y posteriores, seguimos manteniendo el mismo formato que usábamos en las anteriores, en el eje X representamos los distintos grupos que son comparados entre ellos, y, en el eje Y representamos el valor obtenido de esta métrica.

	General	Male	Female	Junior	Senior	Male Junior	Male Senior	Female Junior	Female Senior
EER	3,5397	3,125	2,1875	1,25	2,8125	3,125	10	2,5	6,25
MinDCF	0,23203	0,19062	0,1375	0,09375	0,20313	0,03125	0,2375	0,175	0,125

Tabla 4-5: EER y MinDCF – FairVoice

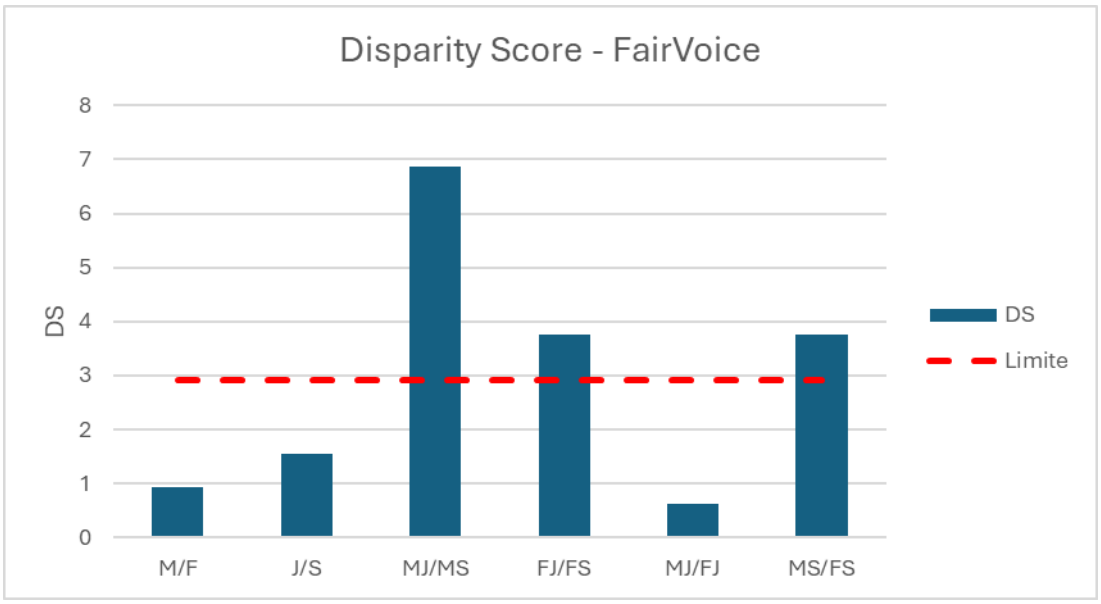

Figura 4-14: *Disparity Score* – FairVoice

En la gráfica de la Figura 4-14, observamos como al evaluar la disparidad de resultado entre los grupos, superan la media la comparación entre los grupos que presentan el mismo género y distintas edades (*male senior / male junior*, y, *female senior / female junior)*, y, entre el grupo que presenta la misma edad y géneros diferentes (*male senior / female senior)*. Especialmente, la disparidad más significativa se está produciendo en el grupo MJ/MS (*male senior / male junior).*

Como hemos visto en los anteriores experimentos, esta disparidad puede estar dada por la base de datos o por el propio sistema. Para ello, vamos a realizar el cálculo de la pérdida de calibración en los datos para descartar o confirmar que el problema viene dado por estos datos.

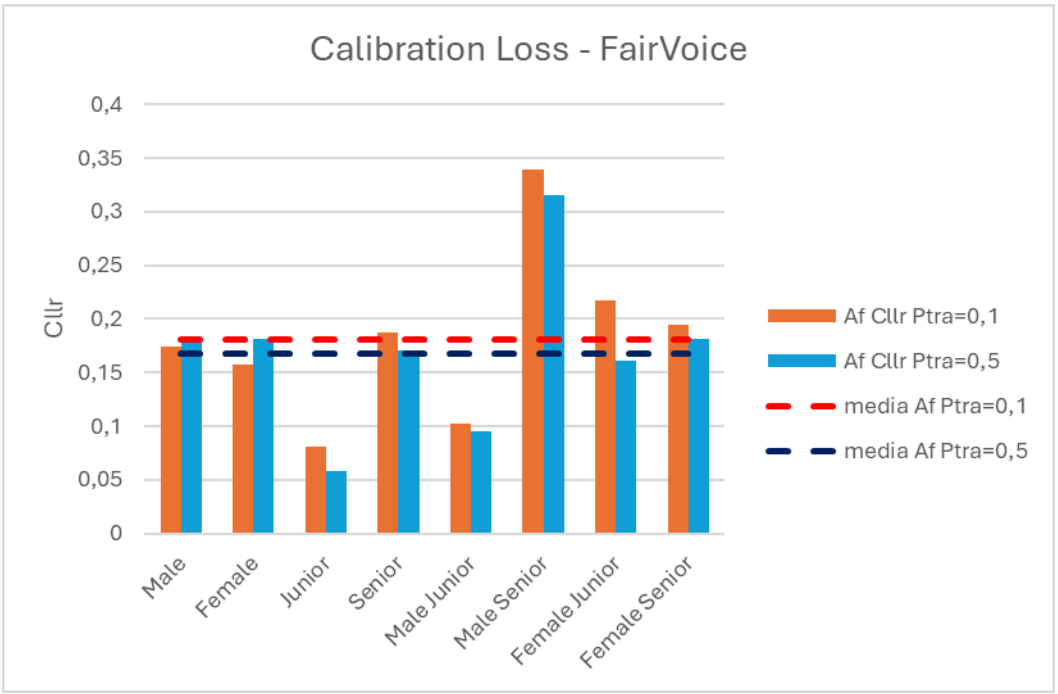

Figura 4-15: *Calibration Loss* – FairVoice

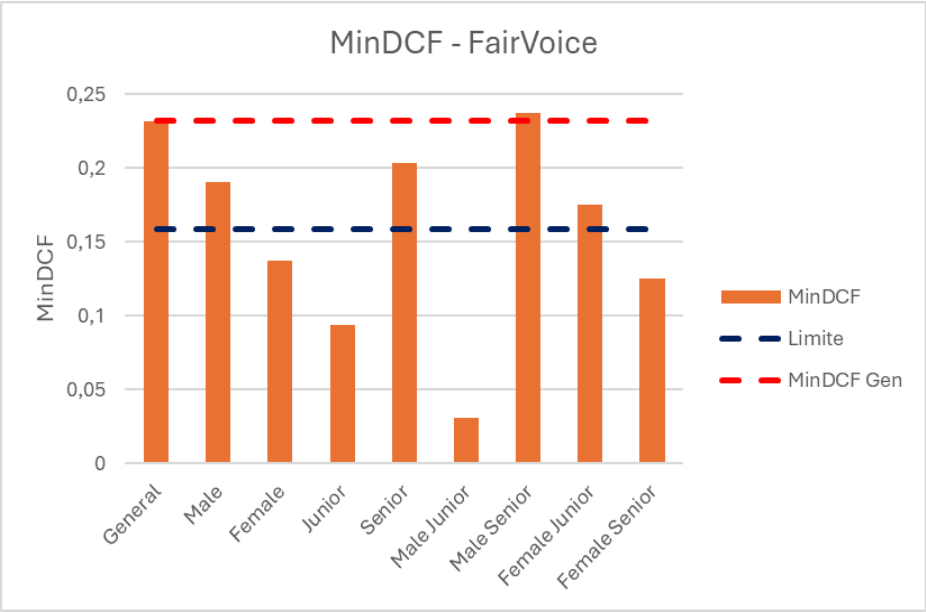

Figura 4-16: MinDCF FairVoice

En la Figura 4-15 se muestra como el grupo *Male Senior* y el grupo *Female Junior*, son los dos grupos que superan la media de la perdida de calibración de todos los grupos. Analizando también el valor mínimo de la función de coste, Figura 4-16, se observa como ambos grupos también superan la media de todos los valores, pero no superan el valor de la evaluación general.

Con esta información, teorizamos que el problema sigue viniendo de los datos, a pesar de contar con la misma cantidad de locutores en las listas, y, la misma cantidad de muestras, en los datos, estos dos grupos se encuentra desequilibrados, generando un sesgo que se refleja también en el sistema.

	Canada	England	Us
EER	2,5	3	3,571
MinDCF	0,025	0,155	0,282

Tabla 4-6: EER y MinDCF – acentos ingleses

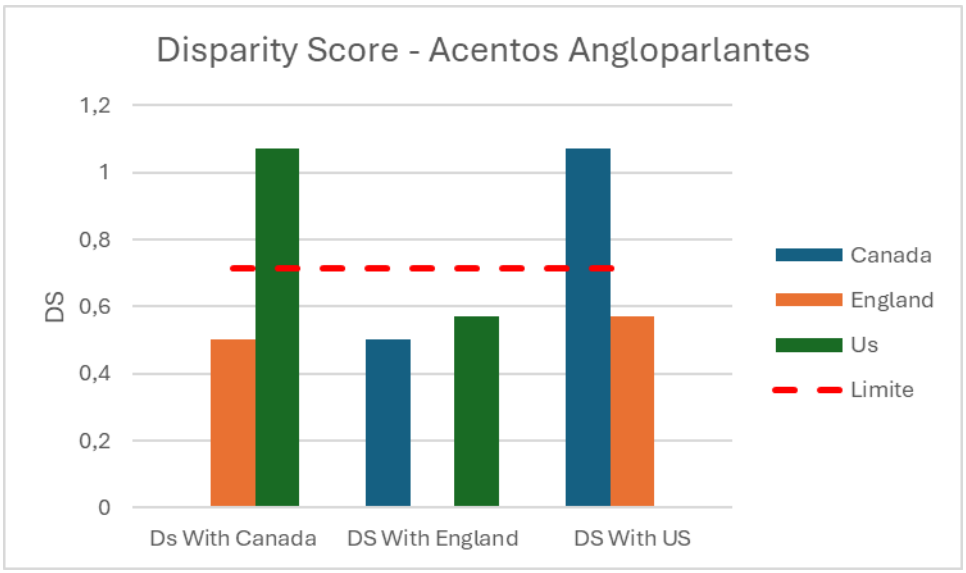

Figura 4-17: *Disparity Score* – acentos angloparlantes

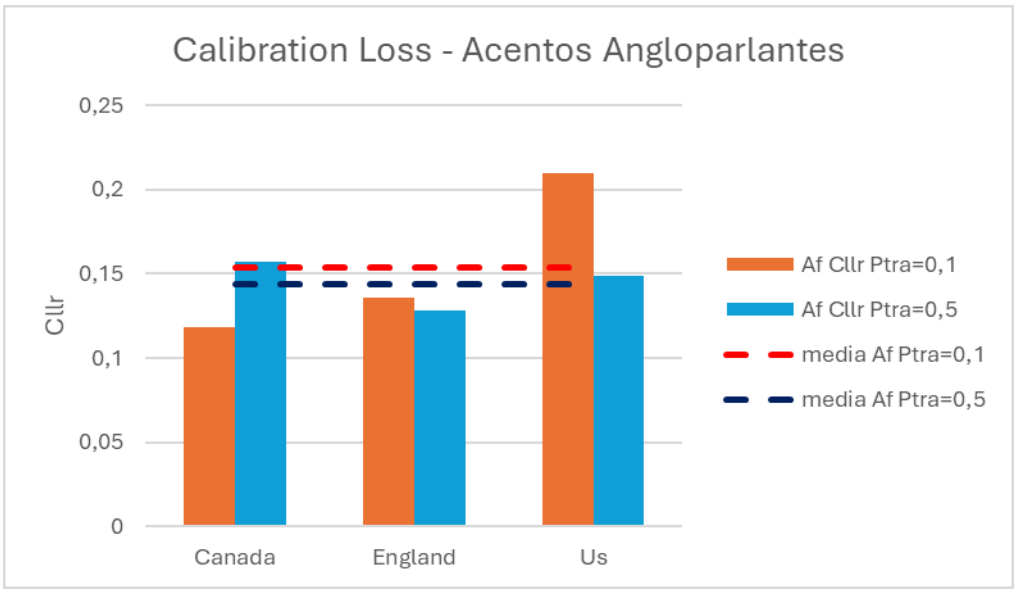

Figura 4-18: *Calibraton Loss* – Acentos ingleses

En la evaluación de la disparidad del resultado en los acentos angloparlantes, se observa en la Figura 4-17 como el grupo de locutores con acento estadounidense (*US*) son sobre los que se está presentando una falta de imparcialidad. En la Figura 4-18 se ilustra como esta falta de imparcialidad se está produciendo por una falta de calibración en los datos, además, analizando la Tabla 4-6, es este grupo quien presentan la mayor tasa de error (EER) y un valor mínimo de coste más alto, en comparación con los demás subconjuntos.

	CaBalearic	CaCentral	CaOther	Es Norte Peninsular
EER	15	6,0714	5	2,5
MinDCF	0,5	0,27857	0,05	0,175

Tabla 4-7: EER y MinDCF – acentos españoles

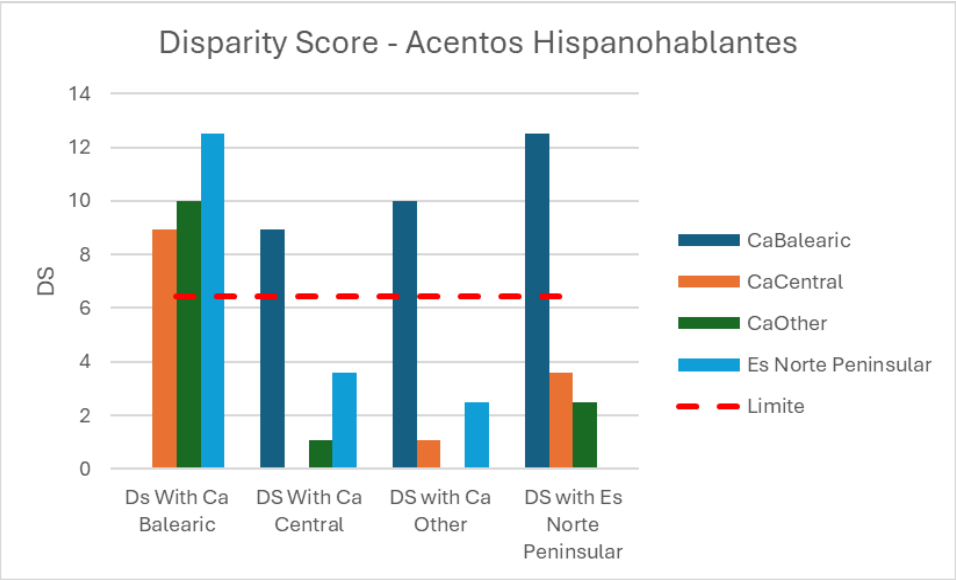

Figura 4-19: Disparity Score – acentos hispanohablantes

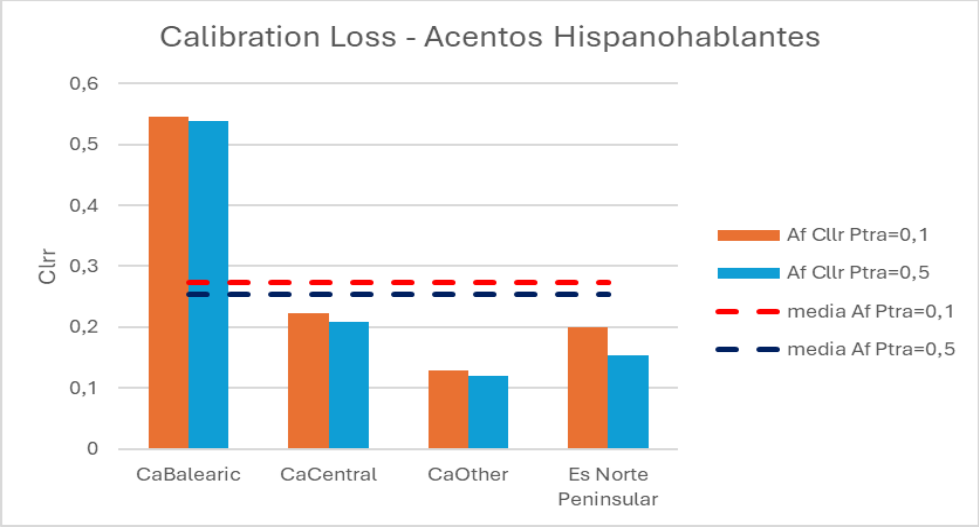

Figura 4-20: Calibraton Loss – acentos hispanohablantes

Para el caso de los acentos hispanohablantes, sucede algo similar que con el grupo de acentos angloparlantes. En este caso, el acento *ca_balearic*, el cual es que presenta los valores de EER y MinDCF (Tabla 4-7) más altos, es, también el que esta generado los valores de disparidad de resultado más altos (Figura 4-19), y, esta imparcialidad que se está dando en los resultados, también vemos en la Figura 4-20 que es justo este acento, el que presenta una pérdida de calibración significativamente por encima de la media.

Con estos resultados en dos conjuntos de acentos, volvemos a plantear las mismas hipótesis que hemos planteado en los anteriores experimentos:

- La poca representación de un grupo en la base de datos puede generar que haya problemas para calibrar los datos, y esto, se acaben afectando al sistema, el cual generara un sesgo para estos grupos y como resultado obtendremos unas tasas de error que presentan parcialidad.

- La teoría que planteábamos en el 4.2 EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE: acentos más generalizados o neutros pueden llevar a que se produzca un sesgo desfavorable para estos locutores, aunque la representación de estos acentos en la base de datos sea alta.

4.4 Experimento 4: Generación de un modelo usando Fine-Tuning para crear una posible solución

Este experimento va a ser similar al experimento 4 que se llevó a cabo en el trabajo de fin de grado [1]. En ese experimento lo que realizamos fue utilizar uno de los modelos ya entrenados del estado del arte [25][27] para sacar los parámetros, y, entrenar un nuevo modelo basados en esos parámetros, pero con el conjunto balanceado de datos de FairVoice.

Los resultados de dicho experimento mostraron que el utilizar uno de los modelos (ya entrenado) del estado del arte, al ser más robusto, mejoraba la tasa de error que presentaban los datos. Pero, este modelo base estaba entrenado con una base de datos desbalanceada, lo que acaba haciendo que el modelo generado con Fine-Tuning, a pesar de ser entrando con una base balanceada, presentaba los sesgos que había en el estado del arte.

La técnica de Fine-Tuning consiste en usar un modelo ya entrenado, generalmente con una base de datos extensa y robusta, para entrenar un nuevo modelo adaptándolo a los datos que tenemos (generalmente una cantidad más reducida). Una de las técnicas es tomar los parámetros del modelo base de algunas de las capas y durante el proceso mantenemos que esos parámetros no se actualicen, entrenando las capas finales.

Lo que vamos a hacer en el este cuarto y último experimento es similar al del trabajo anterior [1], pero en lugar de empelar un modelo base del estado del arte o enterando con una cantidad mayor de datos, lo que hacemos es coger el modelo entrenado con la base de datos balanceada, EXPERIMENTO 3: GENERACIÓN Y EVALUACIÓN DE UN MODELO BALANCEADO CON FAIRVOICE, y, entrenamos las capas finales con las listas del segundo experimento, EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE.

	General	Male	Female	Junior	Senior	Male Junior	Male Senior	Female Junior	Female Senior
EER	4,075	4,057	4,329	3,925	3,946	3,883	3,946	3,846	4,305
MinDCF	0,160	0,145	0,176	0,155	0,176	0,131	0,165	0,120	0,175

Tabla 4-8: EER y MinDCF FairVoice

Para comparar si este experimento reduce la tasa de disparidad que encontramos en EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON

FAIRVOICE**.** Lo que vamos a hacer es comparar en ambos grupos los resultados obtenidos, tal y como pintamos en la Figura 4-21.

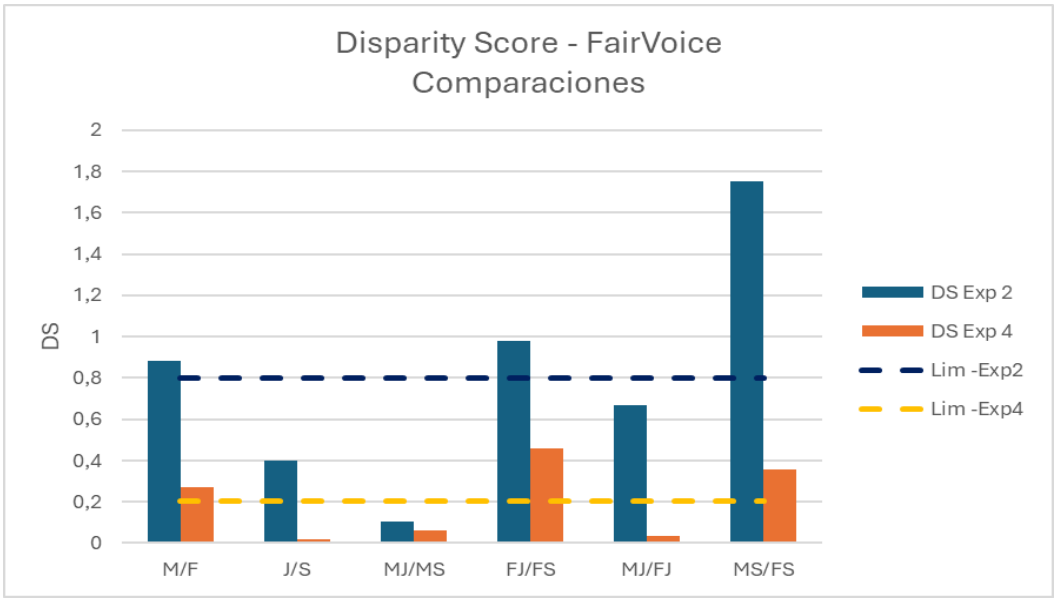

Figura 4-21: Comparación de las disparidades de resultado entre experimentos

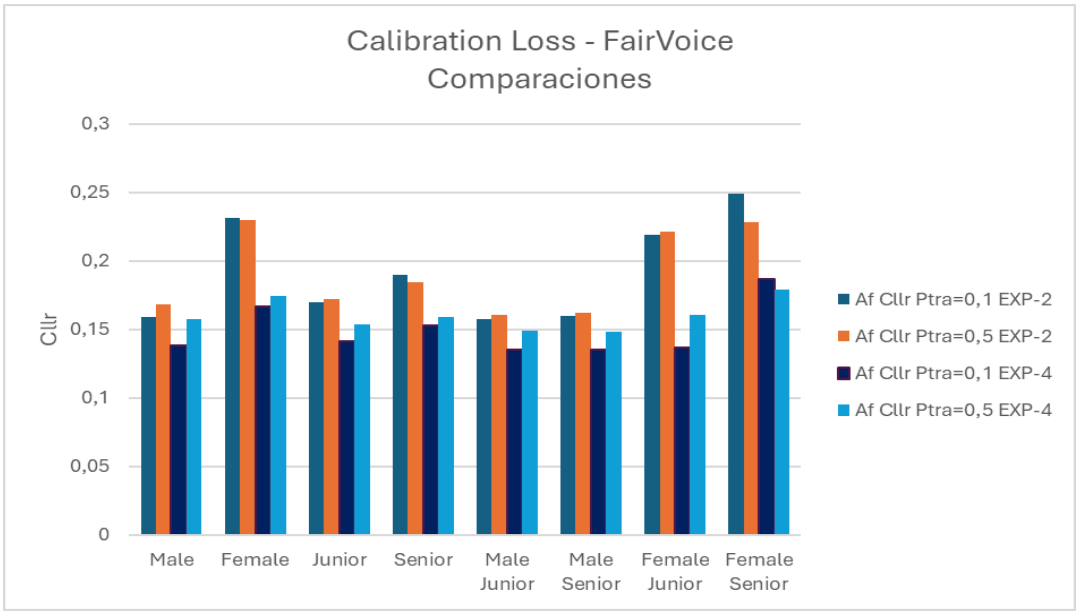

Figura 4-22: Comparación *Calibration Loss* entre experimentos

Como se refleja en la Figura 4-21, en los grupos sesgados por género, edad y la combinación de los dos anteriores, la disparidad se ha reducido entre los grupos que estamos evaluamos, logrando el objetivo de este experimento. Además, al comprar las tasas de error obtenidas (Tabla 4-2, Tabla 4-8 y Figura 4-23), se advierte como en algunos de los subgrupos sesgados (*Female, Senior, Female Junior y Female Senior*) se ha reducido de forma notable la tasa de error. Cuando realizamos la calibración de estos grupos, y comparamos los resultados con los obtenidos con el segundo experimento,

podemos notar una mejora incluso de estos parámetros, tal y como refleja la gráfica de la Figura 4-22.

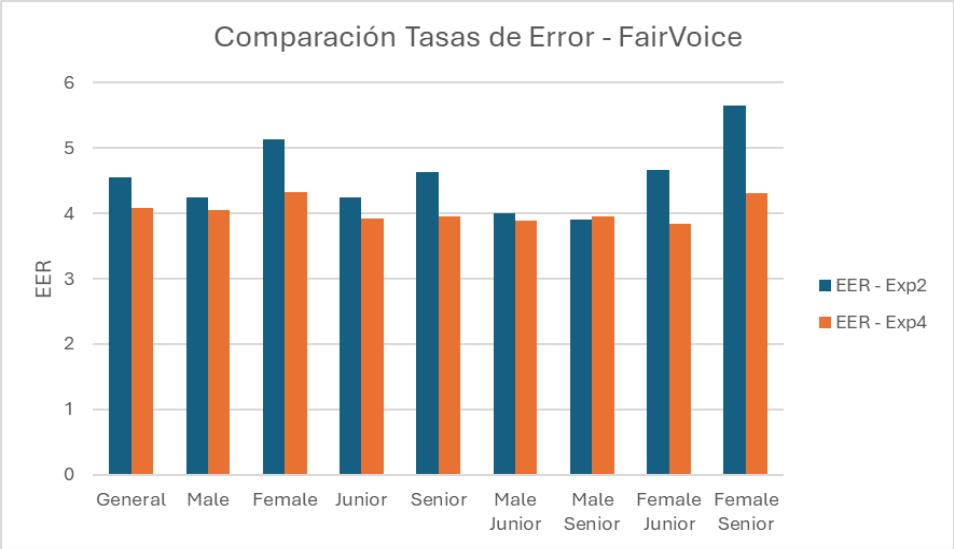

Figura 4-23: Tasas de Error – comparación Exp2 y Exp4

A pesar de los resultados satisfactorios de la primera evaluación de sistema, cuando empleamos las listas de los subgrupos distribuidos en función de los distintos acentos, hemos detectado un error en los resultados con algunos de los subgrupos de acentos tanto angloparlantes como hispanohablantes, y, por este motivo no mostraremos el resultado en la Tabla 4-9 y la Tabla 4-10. Esto motiva a que no hagamos una comparación con los resultados obtenidos en EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE.

	England	Ireland	Us
EER	0,870	0,719	0,870
MinDCF	0,009	0,065	0,017

Tabla 4-9: EER y MinDCF acentos ingleses

	es rio platense	ca Other
EER	1,818	1,250
MinDCF	0,055	0,025

Tabla 4-10: EER y MinDCF acentos españoles

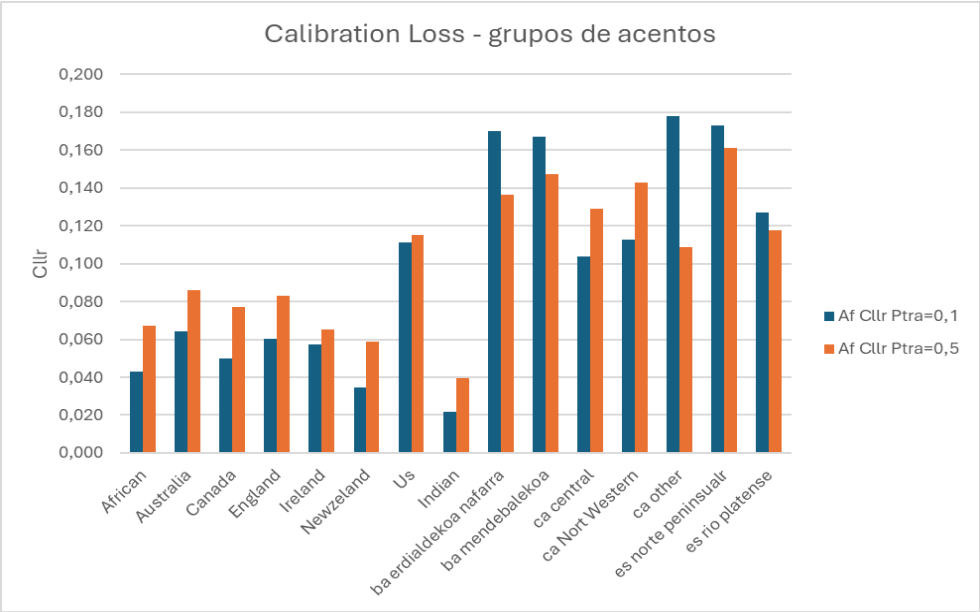

Figura 4-24: *Calibration Loss* – subgrupos de acentos

Con los resultados obtenidos, entendemos que se ha producido un error con el modelo, pero no logramos entender el motivo de este error, ya que se ha empleado la misma estructura que en los anteriores experimentos, y, las mismas listas que en el EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE se empleaban. Al emplear en ellos el código de calibración los datos (Figura 4-24) nos indican que el error es muy posible que se esté dando por algún tipo de problema que no sabemos identificar en el modelo.

Aunque los resultados para los grupos de sesgados por acentos no han sido satisfactorios, y, no entendemos porque se están dando, consideramos que al menos para los grupos sesgados por género y edad, los resultados son satisfactorios. Para estos subgrupos, podemos concluir con los resultados, que emplear de esta forma tan poco convencional la técnica de Fine-Tuning (recordando que generalmente este método emplea un modelo base entrenado con una cantidad significativamente grande de datos, y nuestro caso no se está dando así), ha resultado en una disminución significativa de la disparidad de resultado entre los grupos sesgados por las características privadas.

La conclusión que llegamos con este experimento es que, experimentos que buscan generar primero un modelo o modelos adaptados a unos datos (balanceado, específico para cada género, edad, idioma, etc.) y, posteriormente utilizar técnicas como el Fine-Tuning o similares para adaptarlo a una base de datos real (generalmente estas no se encuentran balanceadas), resulta en una línea de investigación interesante.

Esta línea que mencionamos, puede que en futuras investigaciones llegue al método donde, ya no solo se consiga paliar de forma significativa la falta de imparcialidad en el sistema (como se ha dado en este experimento), sino que también se logren reducir las tasas de error que se estaban presentando, y, tal vez solucionar el problema de los grupos divididos por acento que se están dando en este último modelo.

5 Conclusiones y trabajos futuros

Durante el desarrollo de este trabajo de fin de máster, hemos llevado a cabo distintos experimentos (presentados en el CAPÍTULO 4) con la finalidad de entender porque se producen las faltas de imparcialidad dentro de los sistemas de verificación de locutor cuando estos mismos se dividen en función de sus características privadas (acento, edad y género).

En el primer experimento, EXPERIMENTO 1: EVALUACIÓN DEL ESTADO DEL ARTE CON VOXCELEB, sesgamos a los locutores en función de su género y nacionalidad (como ya se explica en su apartado, esto se toma como acento debido a que todos los clips de audios están en inglés), y, los evaluamos con uno de los modelos pre-entreandos que forma parte del estado del arte en sistemas de verificación de locutor.

Con los datos obtenidos del sistema, calculamos el *disparity score* entre los grupos, y, también realizamos el cálculo de la calibración de los datos de cada subgrupo. El resultado del cálculo del *disparity score* indica que dos de los grupos de acento sobrepasan la media de estos valores, y, cuando realizamos la perdida de calibración, se confirma, que este problema de imparcialidad en estos grupos proviene de una falta de calibración.

Como se explicaba en el CAPÍTULO 3, la base de datos de VoxCeleb se encuentra desbalanceada, especialmente entre las distintas nacionalidades, por lo que planteamos la primera hipótesis: *una base de datos desbalanceada genera problemas de calibración entre los grupos, que, acaban produciendo que el sistema acabe generando sesgos*.

En el segundo experimento, EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE, se realiza de forma similar al primero, pero con la base de datos de FairVoice. La finalidad de este experimento es similar al del anterior, buscamos estudiar los posibles sesgos que se generan con las siguientes características privadas: género, edad y acento, dentro de un modelo entrenado con un base de datos desbalanceada. Y, además, nos va a servir de referencia para el último experimento.

Con los datos obtenidos con los grupos de los locutores divididos en género y edad, volvemos a plantear la misma hipótesis que plateábamos en el anterior experimento, *una base de datos desbalanceada genera problemas de calibración entre los grupos, que, acaban produciendo que el sistema acabe generando sesgos*. Demostrando esta teoría con los resultados, tanto de la disparidad de resultados como con la pérdida de calibración.

Por otro lado, planteamos una nueva hipótesis basada en los acentos, ya que sus resultados no siguen, en todos los casos, la idea general de que cuantos más datos y locutores tengamos con un tipo de característica (en este caso el acento), obtenemos una tasa de error menor. No solo observamos que la cantidad de locutores y datos afecta al resultado, sino que también planteamos la segunda hipótesis: *contar con un acento específico de una región concreta, y, sea un acento más identificativo, puede generar un sesgo favorable frente a acentos más generalizados o neutros*.

En el tercer experimento, EXPERIMENTO 3: GENERACIÓN Y EVALUACIÓN DE UN MODELO BALANCEADO CON FAIRVOICE, evaluamos un modelo generado con la misma base

de datos de FairVoice, pero empleando listas balanceadas en cuestión de locutores y muestras por locutor tanto en entrenamiento como en evaluación. Con este experimento, al tener menos datos, la tasa de error es mayor que en los otros, y, no tenemos una forma de comparar si realizar esta técnica es una mejora en cuestión de las disparidades. Pero, con los resultados obtenidos, volvemos a plantear las hipótesis que habíamos planteado en el anterior experimento.

Por último, en el cuarto y último experimento de este trabajo de fin de máster, EXPERIMENTO 4: GENERACIÓN DE UN MODELO USANDO FINE-TUNING PARA CREAR UNA POSIBLE SOLUCIÓN, lo que hacemos es generar un modelo mediante la técnica de Fine-Tuning, empleando como modelo base, el modelo generado en el tercer experimento, y, usando las listas del segundo experimento para la creación de este nuevo modelo. A pesar de darse un error del sistema con las listas de acentos; conseguimos unos resultados favorables al compararlo con los resultados del modelo del EXPERIMENTO 2: GENERACIÓN Y EVALUACIÓN DE UN MODELO NO BALANCEADO CON FAIRVOICE.

Estos resultados, demuestran que el uso de un modelo balanceado como modelo base para generar uno nuevo, consigue reducir significativamente la disparidad entre los grupos, incluso, reduciendo la pérdida de calibración que se estaba dando. Planteando una solución ante el problema de sesgos que se producían en función del género, la edad y la combinación de ambos. Demostrando que la hipótesis: *una base de datos desbalanceada genera problemas de calibración entre los grupos, que, acaban produciendo que el sistema acabe generando sesgos*, en parte es correcta, y, abriendo una línea de investigación para futuros trabajos.

En este trabajo, hemos analizado los problemas de *fairness* y sus posibles causas, además de propone una posible solución ante dicho problema. Aunque no tengamos un patrón igual en todos los experimentos en función de una característica concreta, sí se da el caso de que hay un patrón. Este patrón, si los grupos cuentan con una desigualdad en su representación en la base de datos, esta va a reflejar esas mismas desigualdades en los resultados, siendo más o menos significativa.

La línea de futuros trabajos que mencionamos previamente, la planteamos como una búsqueda de mejorar las tasas de error que se generan, cuando hemos logrado reducir la disparidad de resultados. Planteamos que se puedan generar nuevos métodos para reducir el problema de *fairness* mediante técnicas que busquen generar modelos adaptados a los datos de forma equitativa o centralizados en datos concretos, para posteriormente generar un modelo con bases de datos más grandes y que no son balanceadas.

Referencias

[1] Aguilera Sepúlveda, A. (2022). Análisis de sesgos en sistemas de reconocimiento de locutor basados en DNN-Embeddings [Trabajo de fin de grado, Universidad Autónoma de Madrid].

[2] Estevez, M., & Ferrer, L. (2022). Study on the Fairness of Speaker Verification Systems on Underrepresented Accents in English. ArXiv, abs/2204.12649.

[3] Hansen, John & Hasan, Taufiq. (2015). Speaker Recognition by Machines and Humans: A tutorial review. Signal Processing Magazine, IEEE. 32. 74-99. 10.1109/MSP.2015.2462851.

[4] S. Misra, T. Das, P. Saha, U. Baruah and R. H. Laskar. (2015). "Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis," 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], Nagercoil, India, 2015, pp. 1-4, doi: 10.1109/ICCPCT.2015.7159307

[5] Reynolds, Douglas & Quatieri, Thomas & Dunn, Robert. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. 10. 19-41. 10.1006/dspr.1999.0361.

[6] Campbell, J. P., Jr. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), 1437-1462. https://doi.org/10.1109/5.628714.

[7] Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 52(1), 12-40. https://doi.org/10.1016/j.specom.2009.08.009

[8] P. Kenny (June, 2010). Bayesian speaker verification with heavy–tailed priors. Keynote presentation, Proc. of Odyssey 2010.

[9] Borgstrom, B.J. (2020). Discriminative Training of PLDA for Speaker Verification with X-vectors.

[10] Ramoji, Shreyas & Krishnan, V & Singh, Prachi & Ganapathy, Sriram. (2020). Pairwise Discriminative Neural PLDA for Speaker Verification.

[11] RedUSERS (Mayo, 2024). Deep learning ¡Crea tu red neuronal! https://www.redusers.com/noticias/publicaciones/crea-tu-red-neuronal-con-deep-learning/Deep.

[12] Snyder, David & Garcia-Romero, Daniel & Sell, Gregory & Povey, Daniel & Khudanpur, Sanjeev. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 5329-5333. 10.1109/ICASSP.2018.8461375.

[13] Rozenberg, S., Aronowitz, H., & Hoory, R. (2020). Siamese x-vector reconstruction for domain adapted speaker recognition. arXiv preprint arXiv:2007.14146.

[14] Cao, Y., Berend, D., Tolmach, P., Levy, M., Amit, G., Shabtai, A., Elovici, Y., & Liu, Y. (2020). Fairness Matters - A Data-Driven Framework Towards Fair and High Performing Facial Recognition Systems. (CoRR). https://dblp.org/db/journals/corr/corr2009.html#abs-2009-05283

[15] Oneto, Luca & Donini, Michele & Pontil, Massimiliano. (2019). General Fair Empirical Risk Minimization. 2020 International Joint Conference on Neural Networks (IJCNN) doi:10.1109/IJCNN48605.2020.9206819

[16] Oneto, Luca & Chiappa, Silvia. (2020). Fairness in Machine Learning. Recent Trends in Learning From Data (pp.155-196), doi: 10.1007/978-3-030-43883-8_7

[17] Zafar, Muhammad & Valera, Isabel & Rodriguez, Manuel & Gummadi, Krishna P. & Weller, Adrian. (2017). From Parity to Preference-based Notions of Fairness in Classification. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, December 2017. 35

[18] Simon Caton & Christian Haas. (2020). Fairness in Machine Learning: a survey. arXiv, doi: 10.48550/ARXIV.2010.04053

[19] Fenu, Gianni & Medda, Giacomo & Marras, Mirko & Meloni, Giacomo. (2020). Improving Fairness in Speaker Recognition. In Proceedings of the 2020 European Symposium on Software Engineering (ESSE 2020). Association for Computing Machinery, New York, NY, USA, 129–136, doi: https://doi.org/10.1145/3393822.3432325

[20] Toussaint, Wiebke & Ding, Aaron. (2021). SVEva Fair: A Framework for Evaluating Fairness in Speaker Verification. arXiv, doi: 10.48550/ARXIV.2107.12049

[21] Shen, H., Yang, Y., Sun, G., Langman, R., Han, E., Droppo, J., & Stolcke, A. (2022). Improving Fairness in Speaker Verification via Group-Adapted Fusion Network. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7077-7081.

[22] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee & Icksang Han. (2020). In defence of metric learning for speaker recognition. Proc. Interspeech 2020, 2977-2981, doi: 10.21437/Interspeech.2020-1064
[23] Ferrer, Luciana & Mclaren, Mitchell. (2020). A Speaker Verification Backend for Improved Calibration Performance across Varying Conditions. 372-379. 10.21437/Odyssey.2020-52.
[24] Mingote, Victoria & Miguel, Antonio & Ortega, Alfonso & Lleida, Eduardo. (2021). Log-Likelihood-Ratio Cost Function as Objective Loss for Speaker Verification Systems. 2361-2365. 10.21437/Interspeech.2021-1085.
[25] VoxCeleb. (Marzo, 2024). https://www.robots.ox.ac.uk/~vgg/data/VoxCeleb/vox1.html
[26] Mozilla Common Voice. (Marzo, 2024). https://commonvoice.mozilla.org/es
[27] GitHub, Clovaai. (Marzo, 2024). VoxCeleb_trainer: In defence of metric learning for speaker recognition. GitHub. https://github.com/clovaai/VoxCeleb_trainer
[28] GitHub, Mirkomarras. (Marzo, 2024). mirkomarras/fair-voice: A Python toolbox for fairness analysis in speaker veriification. GitHub. https://github.com/mirkomarras/fair-voice
[29] GitHub, Luferrer. (Marzo, 2024). luferrer/DCA-PLDA: Discriminative Condition-Aware PLDA. GitHub. https://github.com/luferrer/DCA-PLDA
[30] GitHub, AlmuA (Marzo, 2024) - AlmuA/Reduccion-Sesgo-Speaker-Verificaciton: Modelado de embeddings para la reducción de sesgo en sistemas de reconocimiento de locutor. GitHub. https://github.com/AlmuA/Reduccion-Sesgo-Speaker-Verificaciton
[31] González, J., Brookes, D., & Pardo, J. S. (2014). "Comparing talker variability in speech: The effects of speaking style, background noise, and contrastive emphasis." Journal of the Acoustical Society of America, 136(6), 3040-3049.
[32] Dorta, Josefa & Díaz, Chaxiraxi (2014) “Variables prosódicas en la identificación del locutor”. Quaderns de Filologia: Estudis Lingüístics. XIX: 113-133.
[33] Hansen, J. H. L., & Hasan, T. (2015). "Speaker Recognition by Machines and Humans: A tutorial review." IEEE Signal Processing Magazine, 32(6), 74-99.
[34] Fougeron, C., & Smith, C. L. (1999). "French accents: Phonetic and phonological features." Journal of Phonetics, 27(4), 359-364.
[35] Ramírez Verdugo, M. D. (2006). "Prosodic realization of focus in English and Spanish: A comparative study." Language and Speech, 49(4), 505-529.

Glosario

EER	Equal Error Rate
MinDCF	Minimum Of The Detection Cost Function Computation
Cllr	Calibration Loss
DS	Disparity Score
DNN	Deep Neural Network
PDLA	Probabilistic Linear Discriminant Analysis
SV	Speaker Verification
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MFB	Mel-Filter Bank Outputs
GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model Universal Backgorund Model
TV	Total Variability

Anexos

A Manual para replicar los experimentos

En este anexo, vamos a explicar paso por paso como poder replicar los experimentos que hemos llevado a cabo de manera esquemática, siguiendo el mismo orden que en este trabajo.

Paso 1: clonar el repositorio GitHub [30]. En el repositorio contamos con los siguientes archivos:

- Results: carpeta donde guardamos las tablas y graficas que hemos obtenido con los resultados
- Lists creation: una carpeta donde dentro de ella tenemos los códigos para la creación de las listas
- Scr: carpeta donde contenemos todo el código para poder generar el modelo, y, evaluarlo (basado en los repositorios que hemos mencionado en este trabajo)
- Conjunto de archivos bash para poder entrenar los distintos modelos y evaluar todos los experimentos.

Paso 2: instalar las librerías necesarias en un entorno, usar el archivo requirements.txt

Paso 3: Generación de las listas con el código que tenemos en la carpeta Results

Paso 4: Generar y evaluar los distintos modelos, tal y como hemos descrito en CAPÍTULO 4: EXPERIMENTOS, DESARROLLO Y RESULTADOS

Paso 4.1: Para entrenar se han facilitado los archivos bash que permite entrenar los modelos tal y como se hacen mediante las gpus del grupo de investigación

Paso 4.2: Para evaluar los modelos y conseguir los datos seguimos los siguientes pasos

Paso 4.2.1: Evaluar el modelo primero, con el código bash para obtener los distintos EER, MinDCF y guardar los scores y labels de cada grupo evaluado

Paso 4.2.2: Abrir el archivo de CalibrationExp.py y en, sustituir las rutas que se han marcado en el archivo por las de cada usuario, para poder evaluar la calibración de todos los grupos

Para comprobar si se está ejecutando correctamente, evaluar el experimento 1 como se lleva a cabo en este trabajo con el que siguiente usuario pretende replicar. En la carpeta resultados se encuentra un PDF con los resultados en formas de tablas de cada experimento, evaluarlo con el primero nos asegura que se están haciendo correctamente, ya que es el experimento más fácil de replicar al estar usando un modelo ya entrenado.

"... Mis palabras no has escuchado, transmite lo que has aprendido: fuerza, maestría, pero insensatez, debilidad, fracaso también. ***Sí, fracaso, sobre todo, el mejor profesor el fracaso es****"*

Maestro Yoda, Los Últimos Jedi

There's really no secret about our approach. We keep moving forward - opening up new doors and doing new things - because we're curious. ***And curiosity keeps leading us down new paths****.*

Walt Disney