## 2. Análisis de las técnicas de aprendizaje necesarias para dar ssoporte a la interacción por voz

## 2.1 WhisperAI + Word Embeddings de BERT

Para la interacción con el usuario se necesitará entender aquello que dice el usuario por lo que se deberá interpretar el contenido de la waveform producida para el usuario.

El uso de WhisperAI (Radford, A 2023) permite primero de todo, encontrar una representación textual del contenido de la waveform. A lo largo de la historia han habido muchos modelos de transcripción pero ninguno tan potente como WhisperAI. Este modelo ha sido creado y entrenado por el equipo de OpenAI y su entrenamiento consiste de más de 680000 horas de audio. Un entrenamiento con tantas horas de audio es posible gracias a la arquitectura basada en transformers que emplea, ya que permite explotar con mucha mayor facilidad los recursos computacionales empleados como un número elevado de GPUs. De hecho, es la riqueza de estos datos lo que lo convierte en el modelo de transcripción multilenguaje más popular actualmente. Este modelo es también robusto frente a ruido en la Waveform por lo que permitiría perfectamente que el usuario realizase su anotación en un entorno con mucho ruido como en la calle, que es por donde se moverán los usuarios para realizar las diversas anotaciones. WhisperAI también se trata de un modelo muy accesible y cuyo uso en tareas de voz es sencillo y no requiere de un gran conocimiento experto.

A partir del texto generado por este modelo, podremos encontrar su significado mediante (Almeida, F 2019). Estos se tratan de los pesos de una red neuronal ajustada para predecir a partir de una palabra su contexto o a partir del contexto, la propia palabra. Al ajustar una red para hacer este tipo de predicciones, se condensa la información del contexto en una serie de números reales que forman un vector que proyectado en un espacio dimensional nos permite realizar operaciones aritméticas y geométricas.

Respecto a la tarea de comprender aquello que el usuario nos dice, los Word Embeddings resultan muy útiles dado que nos dan el significado de la palabra en un formato de vector numérico. Su obtención es sencilla dado que se pueden emplear modelos pre entrenados como por ejemplo [BERT](https://arxiv.org/abs/1810.04805) mencionado en (Devlin, J, 2018) Este modelo es capaz de tomar en consideración tanto el contexto de detrás de la palabra que tomamos como referencia como las palabras posteriores por lo que se trata de un modelo muy completo para comprender la transcripción del audio dado por el usuario, también se trata de un modelo de los que compone el estado del arte en NLP actualmente. El modelo cuenta con diversas variantes que añaden complejidad permitiendo obtener embeddings con un mayor o un número de componentes y el uso de un modelo u otro depende de la tarea por lo que se deberá considerar con qué modelo de BERT encontramos representaciones vectoriales de las etiquetas que permitan definir regiones suficientemente delimitadas en el espacio y que mantengan su información semántica es decir, etiquetas que se encuentren en contextos muy distintos, ocuparan a su vez, regiones distintas de este espacio.

2.2 Wav2Vec

Wav2Vec mencionado en (Baevski, A. 2020) se trata del modelo de automatic speech recognition estado del arte más usado y se encuentra entre los más precisos del mercado. Sin embargo, su uso se encuentra menos extendido al de WhisperAI debido a que requiere más conocimiento experto para ser operado. Este modelo también ha demostrado ser robusto frente a distorsiones en la Waveform y ruido de fondo por lo que también sería un buen candidato para ser empleado por los usuarios de la aplicación.

Si con el método anterior teníamos como paso intermedio el texto entre la waveform y el vector de significado, con Wav2Vec no encontramos representaciones intermedias, sino que extrae los vectores de significado directamente a partir del audio. Esto se consigue debido a su arquitectura basada en Encoder-Decoder aplicando atención. Este modelo se puede encontrar preentrenado, y sirve como base para muchos modelos debido a la simplicidad con la que se puede hacer fine tuning sobre el modelo para acabar de ajustarlo a la tarea que se presenta. Debido a su polivalencia, también se puede emplear como modelo de traducción multilenguaje.

Como principal desventaja, Wav2Vec necesita grandes conjuntos de datos etiquetados y sin etiquetar para poder funcionar correctamente y alcanzar grandes niveles de precisión.

2.3 Conformer

Mencionado en (Gulati, A. 2020) se trata de una alternativa a Wav2Vec entrenada por el equipo de DeepMind (Google) y se encuentra también entre los modelos que componen el estado del arte en Automatic Speech Recognition. Respecto a Wav2Vec, presenta un bloque en su arquitectura conocido como Conformer que presenta un mecanismo de atención que permite capturar dependencias en el tiempo entre sonidos a más largo plazo. A pesar de que no presenta una precisión tan elevada como la de otros modelos contemporáneos, es sorprendentemente robusto frente a cambios en el entorno permitiendo hacer la transcripción de voz a vector en entornos difíciles con mucho ruido de fondo o con un hablante alejado al micrófono, por lo que se trata de un método atractivo de cara a permitir a los usuarios de la aplicación realizar grabaciones por la calle.

Esta robustez que presenta se debe principalmente al Conformer debido a que captura relaciones entre sonidos con un mecanismo de atención diferente al que se encuentra presente en texto y es más adecuado para tareas de audio.

2.4 HuBERT

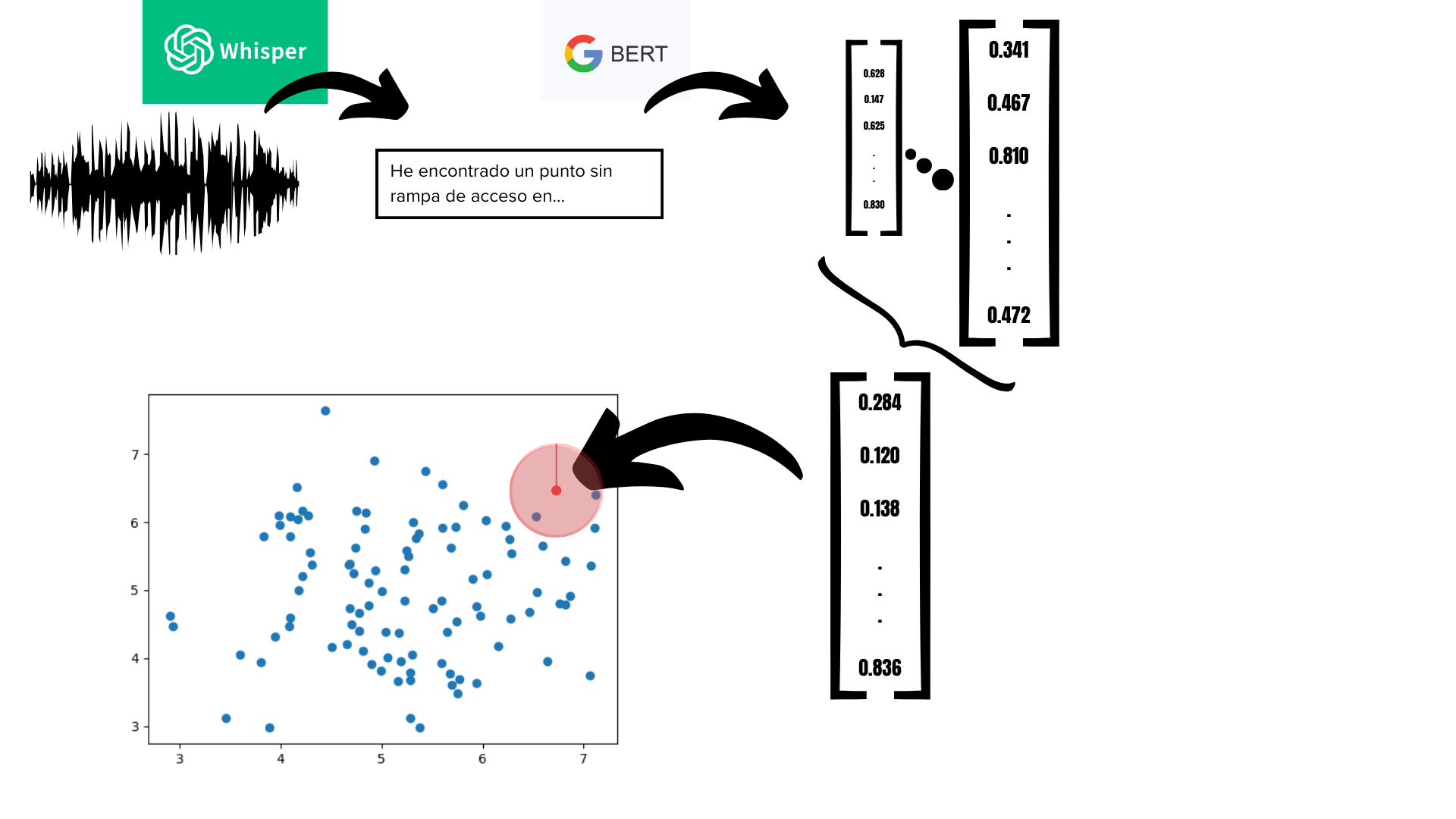
Mencionado en (Hsu, W. 2021), se trata de otra alternativa a Wav2Vec desarrollada por Meta AI que cambia componentes de este como su función de loss, ya que HuBERT emplea la contrastive loss en lugar de la diversity loss, para discretizar la entrada del modelo, no se emplea en este caso Vector Quantization sino un Clustering para discretizar la entrada. En la práctica esto permite clasificar sonidos similares de la misma forma, lo que le da un gran rendimiento al modelo haciéndolo el segundo modelo de Automatic Speech Recognition más potente actualmente siendo el primero el ya mencionado WhisperAI y siendo mejor que éste en tareas de traducción de textos y respuestas a preguntas. Los embeddings extraídos con este modelo son muy útiles para la posterior extracción de tags.

Este modelo es también robusto frente a ruido en la Waveform aunque su robustez depende del proceso de clustering empleado, normalmente se emplea un clustering jerárquico permitiendo este dar resultados muy satisfactorios.

## 3. Análisis de las técnicas de traducción multilenguaje para voz y extracción de “tags”

3.1 Extracción de “TAGS” mediante Radius Nearest Neighbours

Para la extracción de “TAGS” a partir de vectores, siendo esta una tarea de relación de un texto a n tags. Al computar todos los embeddings de un texto, podemos calcular un embedding mediano ponderado según atención a los elementos del texto y a partir del embedding de las diferentes etiquetas, calcular las distancias con todas las etiquetas y mostrar las etiquetas a una distancia menor de un cierto threshold numérico. Este método se llama Radius Nearest Neighbors (RNN) mencionado en (Chen, X. 2022) y se emplea para clustering en diversos escenarios. El problema de este método es una alta sensibilidad a hiper parámetros como la métrica de distancia y el radio que debería tomar en consideración por lo que durante el desarrollo se deberán ajustar meticulosamente estos parámetros. El rendimiento de esta metodología dependerá directamente de la riqueza de los embeddings y si los embeddings de las etiquetas se corresponden a diferentes regiones del espacio de embeddings. Una visualización del proceso completo se puede observar en el siguiente diagrama donde cada punto azul corresponde con una etiqueta y cada punto dentro del radio del embedding promedio se trata de una etiqueta a devolver.





3.2 Extracción de “TAGS” mediante K Nearest Neighbours

Se trata de extraer “TAGS” también a partir de vectores, siguiendo la misma idea de una proyección en un espacio, sin embargo en lugar de considerar el radio, se consideran los K embeddings de etiquetas más próximos al embedding promedio del texto. El principal problema que presenta esta metodología se trata de que ya no encontramos una relación de 1 a n entre el audio y las etiquetas. Sin embargo, nos permite controlar directamente el número de etiquetas que devolvemos como un hiper parámetro. El algoritmo de KNN mencionado en (Cunningham, P. 2020) se trata de un sencillo algoritmo de clustering cuya eficacia se ha comprobado en diversos escenarios y al emplear embeddings como unidades entre las que calcular distancias, también se puede emplear este método ya que está basado en el cálculo de distancias. Lo único que tendríamos que hacer es asegurarnos que las distancias funcionan correctamente con distintos embeddings, ya que la eficacia de KNN depende tanto del hiper parámetro K como de la medida de distancia que se emplee.

3.3 Extracción de “TAGS” mediante HDBSCAN

El uso de un método de clustering por densidades permite devolver multitud de etiquetas similares entre ellas adicionalmente, esto permite definir también una relación de 1 a n entre el texto y las etiquetas. El atractivo de HDBSCAN mencionado en (Malzer, C 2020) como algoritmo de clustering se encuentra en su robustez frente a diversos casos y su sencilla selección de hiper parámetros, ya que a diferencia de otros métodos como DBSCAN no se necesita seleccionar una epsilon ni un mínimo de puntos. Gracias al clustering jerárquico que emplea, se consiguen identificar mejor los clusters que se emplean de embeddings. Una vez computados los clusters, tan solo sería necesario devolver las etiquetas del cluster en el que se encuentra el embedding promedio del texto.

Rehacer el clustering sería necesario a cada vez que hay una nueva entrada de audio y como región de interés tomaríamos tan solo aquella en la que se encuentra el embedding promedio de la transcripción de audio.

3.4 Detección del Idioma

El modelo de detección de idioma desarrollado por SpeechBrain, Voxlingua mencionado en (Ravanelli, M. 2021) presenta varias ventajas significativas que lo hacen idóneo para integrarse en la aplicación de análisis de notas de voz para personas invidentes que navegan por las calles. Principalmente, presenta una precisión excepcional en la detección de idiomas, incluso en condiciones adversas y con una variedad de acentos y dialectos. También se trata de un modelo altamente adaptable y puede entrenarse para reconocer una amplia gama de idiomas, lo que garantiza una experiencia de usuario óptima para personas de diferentes regiones geográficas.

El desafío radica en la detección precisa del idioma hablado en tiempo real. Las soluciones existentes a menudo enfrentan dificultades en este aspecto, lo que puede llevar a errores de interpretación y, en última instancia, a una experiencia de usuario deficiente para las personas invidentes.

Este modelo, ha sido diseñado con un enfoque en la eficiencia computacional, lo que permite su implementación en dispositivos móviles con recursos limitados, lo cual es fundamental para una aplicación destinada a usuarios que utilizan tecnología portátil.

3.5 Traducción de Textos

Se ha decidido emplear el modelo creado por Meta para traducción de textos No Language Left Behind mencionado en (Costa-jussà, M. 2022), este modelo produce traducciones precisas y fluidas, lo que es esencial para garantizar que las personas ciegas comprendan correctamente la información que se les proporciona. Esto es particularmente importante en el contexto de la navegación por la calle, donde la precisión puede ser crucial para la seguridad y la independencia.

Este modelo está entrenado en un conjunto de datos masivo de texto y código, lo que le permite capturar los matices del lenguaje hablado, como las expresiones idiomáticas y el lenguaje informal. Esto es importante para garantizar que las traducciones sean naturales y fáciles de entender para las personas ciegas.

También puede ser entrenado para adaptarse a entornos ruidosos, lo que es importante para las personas ciegas que utilizan aplicaciones de traducción mientras caminan por la calle. Esto puede ayudar a garantizar que las traducciones sean precisas incluso cuando hay ruido de fondo.

En resumen, el modelo No Language Left Behind ofrece varias características que lo convierten en una herramienta valiosa para la traducción de textos extraídos de audios de personas invidentes que navegan por la calle. Su precisión, fluidez, capacidad para capturar matices del lenguaje hablado, adaptación a entornos ruidosos, fácil integración con sistemas de asistencia y potencial para mejorar la calidad de vida lo convierten en una opción atractiva para esta aplicación.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. arXiv preprint arXiv:1901.09069.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Chen, X., & Güttel, S. (2022). Fast and exact fixed-radius neighbor search based on sorting. *arXiv preprint arXiv:2212.07679*.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cunningham, P., & Delany, S. J. (2020). k-Nearest neighbour classifiers: (with Python examples). *arXiv preprint arXiv:2004.04523*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Malzer, C., & Baum, M. (2020, September). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI)* (pp. 223-228). IEEE.

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

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., ... & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.