Grabadora de Voz Inteligente

Juan José Muñoz Noda, Ovido Navarro Pazos y Jesús Armando Padrón

September 25, 2024

Abstract

En este reporte se presenta el desarrollo de una grabadora inteligente capaz de procesar la voz, transcribir el contenido hablado y detectar las emociones de los hablantes utilizando varios modelos preentrenados de aprendizaje automático. La motivación principal del proyecto es mejorar la accesibilidad y el análisis de las interacciones verbales en diversos contextos. Se discuten los modelos utilizados, la arquitectura del sistema, los experimentos realizados y los resultados obtenidos. Además, se abordan las implicaciones éticas del uso de esta tecnología y se sugieren futuras mejoras y aplicaciones potenciales.

Contents

Chapter 1

Introducción

El trabajo con sonido siempre ha sido una tarea desafiante para la computación. Desde los primeros días de la humanidad, el sonido ha jugado un papel crucial en el desarrollo de la cultura, la música y diversos aspectos de la vida en sociedad. Con el avance progresivo de la computación y el desarrollo de hardware más potente, las técnicas de procesamiento de sonido han evolucionado significativamente, logrando resultados cada vez más complejos y precisos.

La computación, desde sus inicios, ha buscado maneras de capturar, analizar y manipular el sonido de formas innovadoras. Inicialmente, estos esfuerzos se centraron en aplicaciones simples, pero con el tiempo, las capacidades tecnológicas han permitido abordar problemas mucho más sofisticados. La mejora en los algoritmos de procesamiento de señal y el incremento en la capacidad de almacenamiento y procesamiento han sido fundamentales para estos avances.

Paralelamente, el desarrollo de la inteligencia artificial (IA) y el aprendizaje automático (ML) ha abierto nuevas fronteras en el procesamiento de sonido. Estas tecnologías han permitido la creación de modelos que pueden aprender y adaptarse a patrones complejos en los datos de audio, facilitando tareas que antes eran inalcanzables.

1.1 Motivación

Gracias a la evolución de las redes neuronales y los modelos de aprendizaje automático (ML), ha surgido la idea de combinar múltiples modelos para crear una grabadora inteligente. Esta grabadora no solo captura audio, sino que también ofrece funcionalidades avanzadas como la identificación del hablante, la detección de emociones en su voz y la transcripción automática de audio a texto.

1.2 Problemática

La parte más importante de este proyecto es lograr integrar todos los modelos utilizados para crear un flujo de los procesos necesitados para el procesamiento de los audios. Tambien es necesario ajustar los modelos externos que son utilizados en estos procesos. La recoleccion de los audios necesarios para el entrenamiento del modelo creado por los integrantes del proyecto, la modificación, procesamiento y etiquetado de los audios.

Chapter 2

Objetivos generales y específicos

2.1 Objetivos Generales

Como objetivos generales se han planteado los siguientes:

1. **La correcta recolección de los audios para los conjuntos de entrenamiento y prueba.** Este objetivo implica la recopilación de muestras de audio de alta calidad, asegurando que se obtenga una diversidad adecuada de voces y entonaciones. Se presta especial atención a la limpieza y preprocesamiento de los datos para garantizar que sean representativos y útiles para el entrenamiento y la evaluación del modelo.

2. **El entrenamiento del modelo para la identificación de los hablantes.** Este objetivo se centra en el desarrollo y ajuste de un modelo de aprendizaje automático capaz de identificar con precisión a diferentes hablantes. Involucra la selección de algoritmos adecuados, la configuración de parámetros óptimos y la realización de múltiples iteraciones de entrenamiento para mejorar la precisión del modelo.

3. **La correcta separación de los hablantes en audios distintos para su posterior procesamiento.** Se busca desarrollar técnicas efectivas para separar las voces de diferentes hablantes cuando están presentes en una misma grabación. Esto incluye el uso de algoritmos de separación de fuentes y técnicas de segmentación de audio para asegurar que cada hablante pueda ser procesado individualmente.

4. **El reentrenamiento de los modelos externos para ajustarlos a las necesidades de este proyecto.** Este objetivo implica tomar modelos preentrenados existentes y ajustarlos para que se adapten mejor a los requisitos específicos del proyecto. Esto incluye la personalización del modelo para mejorar la detección de emociones en el contexto del proyecto y optimizar su rendimiento en las condiciones de audio particulares que se esperan enfrentar.

2.2 Objetivos Específicos

Como objetivos específicos se han planteado los siguientes:

1. Entrenar los modelos para que tengan una precisión lo más acertada posible en las interacciones fuera de los conjuntos de prueba y entrenamiento

2. Que el modelo sea capaz de identificar los hablantes correctamente

3. Lograr crear un flujo de trabajo con todos los modelos utilizados para garantizar la armonía entre estos

4. Devolver con presición el audio transcrito, el nombre del hablante y la emocion predominante en el audio.

Siguiento estos objetivos se ha logrado crear una aplicacion de python en la cual es posible grabar una conversación la cual se procesa y se devuelven los resultados obtenidos de la misma con una presición aceptable y adecuada para este proyecto. Todo esto gracias a los modelos preentrenados y los entrenados por los integrantes del equipo.

Chapter 3

Estado del arte

En esta sección se hace un recuento sobre la literatura acerca de los temas utilizados en este proyecto.

3.1 Separación de Hablantes

PYANNOTE.AUDIO: NEURAL BUILDING BLOCKS FOR SPEAKER DIARIZATION

- **Modelos:** Pyannote.audio es una colección de módulos neuronales entrenables que pueden ser combinados para construir *pipelines* de diarización de locutores. Los modelos incluyen técnicas como *Voice Activity Detection* (VAD), *Speaker Change Detection* (SCD), *Overlapped Speech Detection* (OSD), y *Speaker Embedding* usando *x-vectors*.
- **Dataset:** Los modelos fueron entrenados y evaluados en varios conjuntos de datos, incluyendo:
  - *AMI:* Reuniones.
  - *ETAPE:* Noticias de televisión.
  - *DIHARD:* Diversos dominios con hasta 11 diferentes.
- **Resultados:** Los modelos pre-entrenados alcanzan un rendimiento de última generación en diversas tareas, como la detección de actividad de voz y la detección de cambios de locutor. En pruebas de diarización, el *pipeline* basado en Pyannote.audio mostró una tasa de error de diarización (DER) significativamente reducida en comparación con otros sistemas tradicionales.

Attention-based Encoder-Decoder Network for End-to-End Neural Speaker Diarization with Target Speaker Attractor

- **Modelos:** Este *paper* propone una red de codificador-decodificador basada en atención para la diarización de locutores. El modelo AED-EEND (*End-to-End Neural Diarization*) incorpora información de inscripción del locutor objetivo, utilizando un mecanismo de atención para mejorar el cálculo de los atractores y mitigar el problema de permutación de locutores.

- **Dataset:** El sistema fue entrenado y evaluado en datos simulados para diálogos de 1, 2, 3 y 4 locutores, así como en el conjunto de datos real *CALLHOME* para evaluaciones en grabaciones reales.

- **Resultados:** El sistema AED-EEND superó a los modelos previos como EEND-EDA y TS-VAD en las condiciones de locutores fijos y flexibles, mostrando una menor tasa de error de diarización (DER) especialmente en el conjunto *CALLHOME*.

Powerset multi-class cross entropy loss for neural speaker diarization

- **Modelos:** Este trabajo propone cambiar de una formulación de clasificación multi-etiqueta (usada comúnmente en EEND) a una clasificación multi-clase de *powerset*. Este enfoque asigna clases dedicadas a pares de locutores superpuestos, eliminando la necesidad de un umbral de detección en la etapa de prueba.

- **Dataset:** Se realizaron experimentos en 9 conjuntos de datos de referencia, incluyendo:
  - *AISHELL-4*
  - *AliMeeting*
  - *AMI*
  - *Ego4D*
  - *MSDWild*
  - *REPERE*

- **Resultados:** El enfoque basado en clasificación multi-clase de *powerset* mejoró significativamente el rendimiento de la diarización, particularmente en la detección de superposición de locutores, logrando un rendimiento de última generación en varios de los *benchmarks* mencionados.

3.2 Clasificación de Hablantes

TEXT-INDEPENDENT SPEAKER VERIFICATION USING 3D CONVOLUTIONAL NEURAL NETWORKS

- **Modelos:** Se propone un modelo de redes neuronales convolucionales 3D (3D-CNN) para la verificación de locutores. Este modelo captura información espacial y temporal simultáneamente.

- **Dataset:** Se utiliza el dataset *WVU-Multimodal 2013*, que incluye entrevistas con 1083 locutores en múltiples sesiones.

- **Resultados:** El modelo propuesto mejora significativamente sobre el sistema tradicional basado en *d-vectors*, logrando un 21.1% de EER frente al 24.2% de otros modelos, como el LSTM *End-to-End*.

Text-independent Speaker Verification Using Hybrid Convolutional Neural Networks

- **Modelos:** Se propone un modelo híbrido de redes neuronales convolucionales (combinación de 3D-CNN y 2D-CNN) para la verificación de locutores. Este modelo busca capturar y discriminar la información relacionada con el locutor y la no relacionada simultáneamente.
- **Dataset:** *LibriSpeech*.
- **Resultados:** El modelo híbrido supera los métodos existentes de verificación de locutores, logrando mejoras en la precisión de las tareas de verificación de locutores en escenarios independientes del texto.

3.3 Predecir Emoción del Hablante

STATE-OF-THE-ART SPEECH RECOGNITION WITH SEQUENCE-TO-SEQUENCE MODELS

- **Modelos:** Este *paper* explora mejoras en los modelos de secuencia a secuencia basados en atención, como el modelo *Listen, Attend, and Spell* (LAS). Se implementaron varias mejoras estructurales, incluyendo modelos de sub-palabras (*wordpiece models*) y una arquitectura de atención multi-cabeza (*multi-head attention*). También se exploraron técnicas de optimización como el entrenamiento sincronizado, el *scheduled sampling*, el *label smoothing*, y la optimización de la tasa mínima de error de palabras (*Minimum Word Error Rate* - MWER).
- **Dataset:** Utilizaron un conjunto de datos de aproximadamente 12,500 horas de tráfico de búsqueda por voz de Google, con 15 millones de *utterances* en inglés. El modelo fue evaluado en un conjunto de prueba de aproximadamente 14.8K *utterances* extraídas del tráfico de Google y en un conjunto de 15.7K *utterances* de dictado.
- **Resultados:** Las mejoras en la estructura y optimización del modelo redujeron la Tasa de Error de Palabras (WER) del sistema LAS de 9.2% a 5.6% en tareas de búsqueda por voz, superando a los modelos convencionales, que lograron un WER de 6.7%. En tareas de dictado, el modelo mejorado alcanzó un WER de 4.1%, mientras que el sistema convencional alcanzó un 5%.

Speech Emotion Recognition Based on Deep Residual Shrinkage Network

- **Modelos:** Este trabajo propone un modelo basado en *Deep Residual Shrinkage Network* con *Bidirectional Gated Recurrent Unit* (DRSN-BiGRU). Este modelo combina una red de convolución, una red residual con mecanismos de encogimiento, una unidad recurrente bidireccional, y una red completamente conectada, utilizando un mecanismo de auto-atención para mejorar la capacidad de aprender características efectivas.

- **Dataset:** Se utilizaron tres conjuntos de datos de emociones en el habla:
  - *CASIA:* 9600 pronunciaciones en chino de cuatro oradores con seis emociones (enojo, felicidad, tristeza, miedo, sorpresa y neutralidad).
  - *IEMOCAP:* Más de 12 horas de datos de conversaciones entre actores, con emociones como felicidad, enojo, tristeza y neutralidad.
  - *MELD:* Conjunto de datos derivado de la serie de TV *Friends*, con 13,700 oraciones etiquetadas con siete emociones (enojo, disgusto, tristeza, felicidad, sorpresa, miedo y neutralidad).
- **Resultados:** El modelo DRSN-BiGRU logró una precisión del 86.03% en CASIA, 86.07% en IEMOCAP, y 70.57% en MELD, superando otros modelos comparados como DCNN-LSTM, CNN-BiLSTM y DRN-BiGRU.

3.4 Transcripción de Audio

INVESTIGATING END-TO-END ASR ARCHITECTURES FOR LONG FORM AUDIO TRANSCRIPTION

- **Modelos:** El estudio evalúa tres tipos de modelos ASR (*Reconocimiento Automático de Habla*) de extremo a extremo para transcripción de audio de larga duración:
  - *QuartzNet:* Modelo basado en convoluciones separables de 1D.
  - *ContextNet:* Modelo con convoluciones y módulos de *Squeeze-and-Excitation* (SE) para incorporar contexto global.
  - *Fast Conformer:* Una versión optimizada del Conformer, diseñada para audio de larga duración, que combina atención local con tokens globales.
- **Dataset:** Se evaluaron los modelos en cuatro conjuntos de datos en inglés:
  - *TED-LIUM3:* Conjunto de charlas TED con duración promedio de 16 minutos.
  - *Earnings-21 y Earnings-22:* Llamadas de ganancias en el sector financiero, con audios de hasta 2 horas.
  - *CORAAL:* Entrevistas en inglés con fuerte acento regional y discurso superpuesto.
- **Resultados:**
  - *QuartzNet2* puede procesar audios de hasta 12 horas con una tasa de error de palabras (WER) de 7.31% en TED-LIUM3.
  - *ContextNet* mostró mejor rendimiento que QuartzNet2, con un WER de 5.52% en TED-LIUM3.
  - *Fast Conformer* obtuvo el mejor rendimiento en TED-LIUM3 con un WER de 4.98% cuando se afina con tokens globales. Además, Fast Conformer puede manejar hasta 8 horas de audio en una sola pasada con buena eficiencia.

WhisperX: Time-Accurate Speech Transcription of Long-Form Audio

- **Modelos:** WhisperX es un sistema de transcripción de audio que mejora el modelo Whisper utilizando:
  - *Voice Activity Detection* (VAD): Para segmentar el audio en fragmentos manejables.
  - *Forced Phoneme Alignment*: Para lograr alineación precisa a nivel de palabra.
  - *Cut & Merge*: Para dividir y unir segmentos de audio optimizando la transcripción paralela.
- **Dataset:** Se evaluó WhisperX en varios conjuntos de datos:
  - *TED-LIUM3:* Para transcripción de charlas TED.
  - *Kincaid46:* Videos de YouTube.
  - *AMI Meeting Corpus:* Reuniones grabadas con alineación a nivel de palabra.
  - *Switchboard-1 (SWB):* Conversaciones telefónicas.
- **Resultados:**
  - WhisperX superó a los modelos Whisper y wav2vec2.0 en términos de precisión de segmentación y velocidad de transcripción.
  - WhisperX, utilizando VAD y alineación forzada, alcanzó un WER de 9.7% en TED-LIUM3 con una mejora de la velocidad de transcripción hasta 11.8 veces en comparación con Whisper.
  - En segmentación a nivel de palabra, WhisperX logró una precisión de 84.1% y un *recall* de 60.3% en AMI.

Chapter 4

Propuestas de solución

Para resolver esta problemática es necesario dividir el problema en partes:

4.1 Separación de Hablantes

El primer paso es, del audio que se quiere procesar, lograr separar los hablantes. Esta tarea es crucial, ya que permite el análisis individual de cada voz, facilitando posteriores procesos como la transcripción y la detección de emociones. Para lograr esto, se ha utilizado un modelo preentrenado de la biblioteca de Python `pyannote.audio`. Este modelo ha sido seleccionado debido a su alta precisión en la separación de hablantes en diversas condiciones de grabación. En este paso se separa el audio principal en segmentos de audios los cuales pasaran a ser procesados mas tarde

El modelo de `pyannote.audio` utiliza técnicas avanzadas de procesamiento de señales y aprendizaje profundo para identificar y separar las diferentes voces presentes en una grabación. Sin embargo, la calidad del audio juega un papel fundamental en la efectividad de este modelo. Audios con ruido de fondo, grabaciones de baja calidad o situaciones donde los hablantes se superponen pueden presentar desafíos significativos. En tales casos, el modelo puede cometer errores, como confundir las voces o no separar adecuadamente a los hablantes.

4.2 Identificar Hablante

Después de separar los hablantes, y asumiendo que dicha separación ha sido efectiva, el siguiente paso es identificar a la persona que está hablando. Para lograr esto, se ha creado un conjunto de audios de los integrantes del equipo, los cuales fueron utilizados para entrenar y evaluar el modelo de identificación de hablantes.

Este conjunto de datos se ha dividido en dos subconjuntos: uno de entrenamiento y otro de prueba, siguiendo una proporción del 70 y 30 respectivamente. Esta división permite entrenar el modelo con una cantidad significativa de datos mientras se reserva una porción para evaluar su desempeño y generalización a nuevos datos no vistos durante el entrenamiento.

El modelo de identificación de hablantes fue desarrollado desde cero por los integrantes del equipo, empleando técnicas de aprendizaje supervisado. Se utilizaron diversas bibliotecas de Python especializadas en la creación y entrenamiento de redes neuronales, como TensorFlow y scikit-learn (sklearn). Estas herramientas proporcionaron los métodos

necesarios para construir y optimizar un modelo de aprendizaje profundo capaz de reconocer a los hablantes con alta precisión.

El proceso de desarrollo del modelo incluyó varias etapas clave:

1. **Preprocesamiento de Datos**: Los audios recopilados fueron normalizados y preprocesados para asegurar una calidad consistente y eliminar ruido. Esto incluye técnicas de reducción de ruido y normalización del volumen.

2. **Extracción de Características**: Se extrajeron características relevantes del audio, como coeficientes cepstrales de frecuencia mel (MFCCs), que capturan la información esencial para diferenciar las voces de los hablantes.

3. **Construcción del Modelo**: Se diseñó una red neuronal utilizando TensorFlow, adaptada específicamente para la tarea de identificación de hablantes, donde los parametros de entrada son las caracteristicas extraidas previamente. La arquitectura del modelo se optimizó para balancear precisión y eficiencia computacional.

4. **Entrenamiento del Modelo**: Utilizando el subconjunto de entrenamiento, el modelo fue entrenado iterativamente, ajustando sus parámetros para minimizar el error de predicción.

5. **Evaluación y Validación**: El subconjunto de prueba se utilizó para evaluar el desempeño del modelo, asegurando que pueda generalizar bien a nuevos datos.

4.3 Identificar Emoción en el Hablante

Al finalizar la identificación del hablante, nuestro objetivo es predecir su estado emocional basándonos en sus grabaciones de audio. Para abordar esta tarea, hemos empleado el modelo Wav2Vec2, una arquitectura avanzada diseñada inicialmente para la transcripción automática de voz. Este modelo, entrenado en una gran cantidad de datos de habla, captura tanto las características acústicas como lingüísticas de manera robusta.

Aplicamos Fine Tuning al Wav2Vec2 para adaptarlo específicamente a nuestro problema de clasificación emocional. Definimos seis etiquetas que representan las emociones clave que deseamos identificar en el hablante, incluyendo tristeza, ira, disgusto, miedo, felicidad y neutralidad. El proceso de Fine Tuning implica ajustar los pesos del modelo en función de nuestro conjunto de datos etiquetado, permitiendo que el modelo aprenda y se especialice en la clasificación precisa de estas emociones.

Para este propósito, utilizamos un conjunto de datos suplementario obtenido de fuentes en línea. Aunque este conjunto de datos está en inglés, su diversidad y calidad han demostrado ser altamente efectivas durante nuestros experimentos. Esta decisión se basa en la capacidad del modelo para generalizar patrones emocionales a partir de diferentes contextos lingüísticos y acústicos, mejorando así la precisión y la capacidad de generalización del modelo en escenarios prácticos.

4.4 Transcripción del audio

Una vez identificado al hablante y evaluado su estado emocional, el siguiente paso es convertir el audio correspondiente en texto. Para este propósito, utilizamos la biblioteca

de Python Google Speech Recognition, la cual se basa en un avanzado servicio de reconocimiento de voz desarrollado por Google. Este servicio permite la conversión instantánea del habla en texto en tiempo real, facilitando la automatización de procesos que requieren transcripción precisa y eficiente.

El proceso implica iterar a través de todos los audios previamente separados y aplicar este servicio para obtener la transcripción de cada archivo de audio individual. Google Speech Recognition emplea modelos y algoritmos optimizados, entrenados con grandes volúmenes de datos, lo que asegura una alta precisión en la conversión, incluso en entornos con diferentes acentos y condiciones acústicas.

Chapter 5

Experimentación y resultados

Como parte del proceso de experimentacion se ha sacado el promedio de las caracteristicas de audio de los integrantes del equipo reflejado en la siguiente imagen. Tener en cuenta que los graficos siguientes se han promediado de un conjunto de mas de 700 audios. Al menos 200 audios por cada integrante:

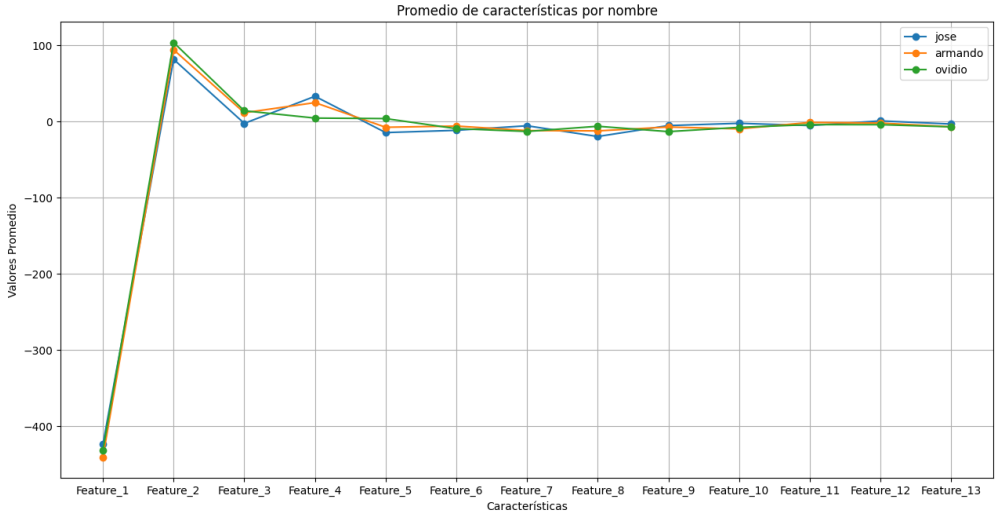

Figure 5.1: Caracteristicas de los hablantes

Mapa de correlacion entre las caracteristicas:

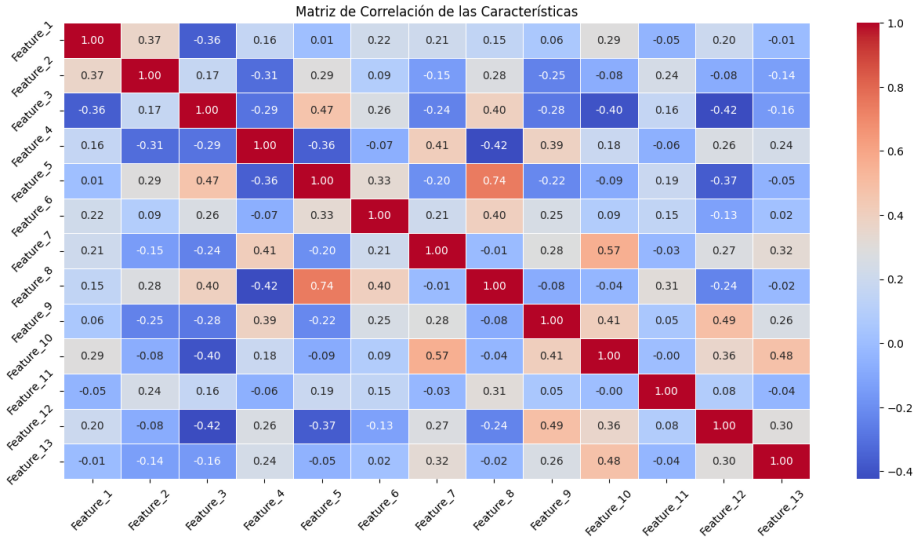

Figure 5.2: Correlación de las características

5.1 Experimentos

Para la realización de los experimentos se han utilizado 10 audios de 1min cada 1 en los cuales intervienen los diferentes integrantes del equipo de forma intermitente:

Audio 1: Aquí el modelo fue capaz de detectar a todos los hablantes con gran precisión sin equivocarse. Además detecta las emociones con una presición similar aunque comete algunos fallos. En la fase de transcripción se desempeña bastante bien, pero confunde algunas palabras, ya que el dialecto utilizado en el habla puede ser complejo

Audio 2: En este audio se puede notar una sucesion de errores como se esperaba ya que los hablantes superponen las voces y el modelo de separacion de los audios comete errores y los demas modelos arrastran este error haciendo predicciones erróneas tanto del hablante como de los sentimientos.

Audio 3: En este audio se exageran las emociones con el objetivo de poner a prueba el modelo encargado de esto. En los resultados se pudo observar que: la separacion de los hablantes fue exitosa, la identificacion de los mismos fue muy precisa exeptuando 2 casos en los que intercambia a los hablantes. La predicción de los sentimientos fue exitosa en todos los casos junto con la transcripción.

Audio 4: Este audio se puso a prueba el modelo de transcripcion, los hablantes dicen sus oraciones con rapidez. La separación de hablantes fue bastante existosa, exeptuando casos en los que los hablantes cambian con rapidez y no logra separa bien el momento exacto del cambio. La clasificación de los hablantes con buena presicion y la transcripcion tiene algunos errores pero todavia se asemeja bastante al audio.

Audio 5: En este, los hablantes actúan de forma pausada y respetando los tiempos del habla de los demás. La separacion, clasificación, sentimiento, transcripción con muy buena presición.

Audio 6: Los hablantes actuan de forma normal, Se aprecia una buena presición en la separación, la clasificación con errores menores, la deteccion sentimental con buena presición y la transcripción bastante bien.

Audio 7: Semejante al experimento 3, pero en este caso se equivoca detectando 1 emoción y 2 hablantes intercambiados. La transcripcion fue exitosa de igual forma

Audio 8: Similar al experimento 1, los hablantes actuan de forma natural y es capaz de reconocerlos a todos con gran precisión, buena identificacion de sentimientos y buena transcripción.

Audio 9: Los hablantes actuan de forma normal. Los diferentes modelos actuan correctamente.

Audio 10: Los hablantes actuan con sigilo como si no quisieran ser escuchados por los demas. El modelo de separacion actua con presición, la clasificacion con algunas deficiencias, la deteccion desentimientos con buena presición y la transcripcion muy buena.

Figure 5.3: Resumen de Experimentos

Chapter 6

Discusión de los resultados

Como se ha podido apreciar gracias a la experimentacion con los diferentes modelos, los resultados que salen a la vista son evidentes

1. **La calidad del audio sí importa:** En los audios donde existe claridad entre los hablantes es más fácil la separacion de los hablantes y la clasificación de los mismos
2. **Escucha lo que dice tu compañero:** Los audios donde los hablantes superponen las voces causan problemas en la separación de hablantes y estos se acarrean a los demás modelos,
3. **Demuestra tus sentimientos:** Cuando en el audio procesado es facil identificar si una persona está feliz o alterada, el modelo de reconocimiento de emociones funciona de maravilla.
4. **Despacio...:** En los audios donde los hablantes van muy deprisa el modelo de transcripcion comete errores.

Estos experimentos no solo revelan las fortalezas del modelo en condiciones controladas y su capacidad para adaptarse a diversos estilos de habla, sino que también destacan áreas críticas para la mejora. Abordar estos desafíos mediante técnicas avanzadas y una diversificación del entrenamiento promete mejorar significativamente el rendimiento del modelo, haciéndolo más robusto y preciso para aplicaciones en el mundo real.

Chapter 7

Conclusiones y trabajo futuro

Tras la finalización de este proyecto, se ha identificado un amplio espectro de posibles aplicaciones tanto en ámbitos científicos como comerciales. Por ejemplo, la grabación de sesiones de psicoterapia para análisis posterior, el registro completo de procedimientos judiciales, el seguimiento de discusiones en grandes reuniones corporativas e incluso la captura de conversaciones informales durante entrevistas. Todos estos usos serían factibles mediante un reentrenamiento de los modelos específicamente adaptado a cada caso, con la inclusión de un nuevo conjunto de datos de entrenamiento que contenga grabaciones de los participantes.

Como trabajo futuro, se podrían integrar nuevas funcionalidades a la grabadora de voz, como el procesamiento de texto para detectar inconsistencias o mentiras en las grabaciones, y para identificar declaraciones contradictorias. Además, sería posible desarrollar una interfaz intuitiva que facilite la manipulación y configuración de la grabadora, mejorando así su accesibilidad para los usuarios finales.

Estas mejoras no solo podrían ampliar significativamente las aplicaciones prácticas de la tecnología desarrollada, sino también aumentar su utilidad en una variedad de contextos profesionales y personales, promoviendo una mayor transparencia y precisión en la captura y análisis de información vocal.

Bibliography

[1] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124-7128.

[2] Chen, Z., Han, B., Wang, S., & Qian, Y. (2023). Attention-based Encoder-Decoder Network for End-to-End Neural Speaker Diarization with Target Speaker Attractor. *ArXiv*, abs/2305.10704.

[3] Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *ArXiv*, abs/2310.13025.

[4] Torfi, A., Nasrabadi, N. M., & Dawson, J. M. (2017). Text-Independent Speaker Verification Using 3D Convolutional Neural Networks. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1-6.

[5] Selin, M., & Mathew, D. P. (2021). Text-independent Speaker Verification Using Hybrid Convolutional Neural Networks. *Webology*, 18, 756-766.

[6] Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2017). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4774-4778.

[7] Han, T., Zhang, Z., Ren, M., Dong, C., Jiang, X., & Zhuang, Q. (2023). Speech Emotion Recognition Based on Deep Residual Shrinkage Network. *Electronics*.

[8] Koluguri, N. R., Kriman, S., Zelenfroind, G., Majumdar, S., Rekesh, D., Noroozi, V., Balam, J., & Ginsburg, B. (2023). Investigating End-to-End ASR Architectures for Long Form Audio Transcription. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 13366-13370.

[9] Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Interspeech*.

[10] Alsabhan, W. (2023). Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. *Sensors (Basel, Switzerland)*, 23.

[11] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv*, abs/2006.11477.