

Grabadora de Voz Inteligente

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

July 8, 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

| | | |
|----------|----------------------------------------------|-----------|
| 1 | Introducción | 2 |
| 1.1 | Motivación | 2 |
| 1.2 | Problemática | 2 |
| 2 | Objetivos generales y específicos | 3 |
| 2.1 | Objetivos Generales | 3 |
| 2.2 | Objetivos Específicos | 3 |
| 3 | Estado del arte | 5 |
| 3.1 | Separación de Hablantes | 5 |
| 3.2 | Clasificación de Hablantes | 6 |
| 3.3 | Predecir Emoción del Hablante | 6 |
| 4 | Propuestas de solución | 8 |
| 4.1 | Separación de Hablantes | 8 |
| 4.2 | Identificar Hablante | 8 |
| 4.3 | Identificar Emoción en el Hablante | 9 |
| 4.4 | Transcripción del audio | 9 |
| 5 | Experimentación y resultados | 11 |
| 5.1 | Experimentos | 12 |
| 6 | Discusión de los resultados | 14 |
| 7 | Conclusiones y trabajo futuro | 15 |

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 necesarios para el procesamiento de los audios. También es necesario ajustar los modelos externos que son utilizados en estos procesos. La recolección 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 precisión el audio transcrito, el nombre del hablante y la emoción predominante en el audio.

Siguiento estos objetivos se ha logrado crear una aplicación 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 precisión 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

La separación de hablantes es un campo activo de investigación en procesamiento de señales de audio y aprendizaje automático, que busca la capacidad de discernir y separar las contribuciones de múltiples hablantes en una mezcla de audio. Aquí tienes un resumen del estado del arte actual en este campo:

Técnicas Tradicionales y Métodos Clásicos

1. **Beamforming y Array Processing:** Utilización de matrices de micrófonos para enfocar la atención en una fuente específica y reducir la interferencia de otros hablantes.
2. **Análisis en el Dominio de la Frecuencia:** Técnicas como la Transformada de Fourier para separar señales de audio basadas en sus componentes frecuenciales.
3. **Algoritmos de Agrupamiento y Modelos de Máxima Verosimilitud:** Métodos que modelan la mezcla de señales de audio como una superposición de fuentes separadas y utilizan técnicas estadísticas para separarlas.

Herramientas y Conjuntos de Datos

1. **Bases de Datos de Evaluación:** Conjuntos de datos como el WSJ0 (Wall Street Journal Zero Resource), CHiME-3 y otros, utilizados para evaluar y comparar algoritmos de separación de hablantes.
2. **Bibliotecas y Frameworks:** Desarrollo de bibliotecas como PyTorch y TensorFlow con módulos específicos para implementar y evaluar algoritmos de separación de hablantes utilizando modelos de aprendizaje profundo.

3.2 Clasificación de Hablantes

La clasificación de hablantes es un área de investigación en procesamiento de señales de audio y aprendizaje automático que se enfoca en identificar y distinguir entre diferentes hablantes basándose en características acústicas y lingüísticas de su voz.

Técnicas Tradicionales y Métodos Clásicos

1. **Modelos Gaussianos Mixtos (GMM):** Utilizados históricamente para modelar y clasificar hablantes basándose en características estadísticas extraídas del espectro de frecuencia del habla.
2. **Vector de Características de Mel-Frecuencia Cepstral (MFCC):** Representación comúnmente utilizada para capturar las características distintivas del habla, que se usa en combinación con clasificadores tradicionales como SVM o k-NN.
3. **Identificación Basada en Características de Voz:** Métodos que utilizan medidas físicas de la voz como la frecuencia fundamental (pitch), la energía y el formato vocal para diferenciar hablantes.

Herramientas y Conjuntos de Datos

1. **Bases de Datos de Evaluación:** Conjuntos de datos como TIMIT, VoxCeleb y LibriSpeech, utilizados para entrenar y evaluar algoritmos de clasificación de hablantes y comparar su rendimiento.
2. **Bibliotecas y Frameworks:** Desarrollo de bibliotecas como TensorFlow y PyTorch que incluyen módulos específicos para implementar y evaluar modelos de clasificación de hablantes utilizando técnicas de aprendizaje profundo.

3.3 Predecir Emoción del Hablante

La predicción de la emoción del hablante es un campo de estudio interdisciplinario que combina técnicas de procesamiento de señales, aprendizaje automático y psicología para identificar y clasificar las emociones expresadas en el habla humana.

Técnicas Tradicionales y Métodos Clásicos

1. **Extracción de Características Acústicas:** Uso de características acústicas como la frecuencia fundamental, la intensidad y el espectro de frecuencia para capturar aspectos emocionales en el habla.
2. **Análisis Prosódico:** Estudio de la entonación, ritmo y duración de las sílabas y frases para inferir estados emocionales.
3. **Modelos Estadísticos:** Aplicación de modelos como Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN) para clasificar emociones basadas en características extraídas del habla.

Herramientas y Conjuntos de Datos

1. **Bases de Datos Anotadas:** Conjuntos de datos como IEMOCAP, EmoDB y MSP-IMPROV que contienen grabaciones de audio etiquetadas con emociones humanas para entrenamiento y evaluación de modelos.
2. **Librerías y Frameworks:** Desarrollo de herramientas y bibliotecas en Python como librosa, TensorFlow y PyTorch que facilitan la extracción de características, entrenamiento de modelos y evaluación de desempeño en la predicción de emociones.

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 características extraídas 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 características 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:

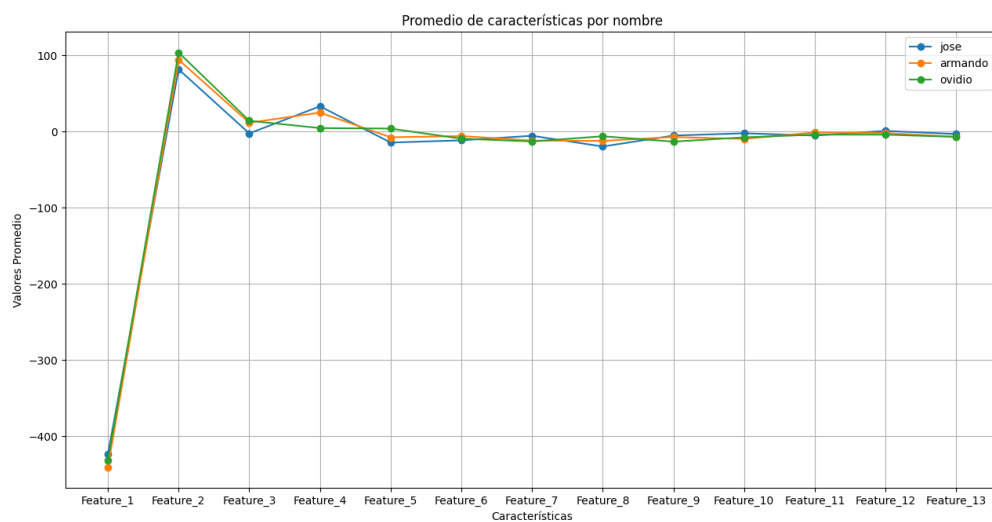


Figure 5.1: Características de los hablantes

Mapa de correlacion entre las características:

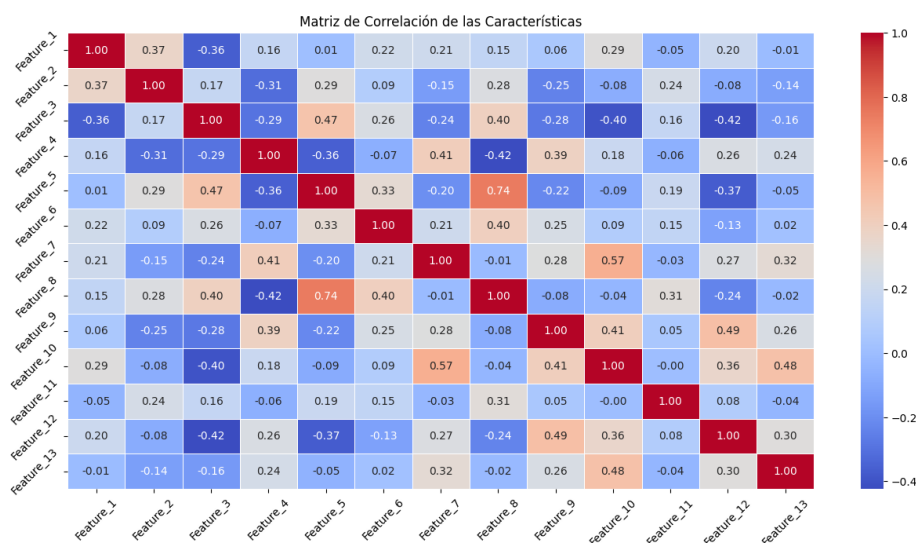


Figure 5.2: Correlación de las características

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 precisión 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 sucesión de errores como se esperaba ya que los hablantes superponen las voces y el modelo de separación de los audios comete errores y los demás 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 separación de los hablantes fue exitosa, la identificación de los mismos fue muy precisa exceptuando 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 transcripción, los hablantes dicen sus oraciones con rapidez. La separación de hablantes fue bastante exitosa, exceptuando casos en los que los hablantes cambian con rapidez y no logra separar bien el momento exacto del cambio. La clasificación de los hablantes con buena precisión y la transcripción tiene algunos errores pero todavía 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 separación, clasificación, sentimiento, transcripción con muy buena precisión.

Audio 6: Los hablantes actúan de forma normal, Se aprecia una buena precisión en la separación, la clasificación con errores menores, la detección sentimental con buena precisión 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 transcripción fue exitosa de igual forma

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

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

Audio 10: Los hablantes actúan con sigilo como si no quisieran ser escuchados por los demás. El modelo de separación actúa con precisión, la clasificación con algunas deficiencias, la detección de sentimientos con buena precisión y la transcripción muy buena.

Chapter 6

Discusión de los resultados

Como se ha podido apreciar gracias a la experimentación 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 separación 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 fácil 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 transcripción 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] M. Selin, Dr. K. Preetha Mathew *Text-independent Speaker Verification Using Hybrid Convolutional Neural Networks*, 2023.
- [2] Tian Han ^{1,2}, Zhu Zhang ^{1,2,*}, Mingyuan Ren ¹, Changchun Dong ¹, Xiaolin Jiang ¹ and Quansheng Zhuang, *Speech Emotion Recognition Based on Deep Residual Shrinkage Network*, electronics, 2023.
- [3] Waleed Alsabhan, *Human-Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention*, sensors, 2023.
- [4] *pyannote.audio: neural building blocks for speaker diarization*, <https://arxiv.org/abs/1911.01255v1>