Aprendizaje Profundo Para Procesamiento de Audio

Practica 1 - Detección de eventos de audio en DCASE

Universidad Autónoma de Madrid

Escuela Politécnica Superior

Master en Ciencia de Datos 2021/2022

Guillermo Hoyo Bravo

1.- Preparación de entorno:

A) Entonrno Software

Hecho

B) Descripción de la base de datos

2.- Representación de Segmentos de audio y las anotaciones

a) Formas de onda y características de audio

Elija un fichero de audio cualquiera en la carpeta dataset/audio/validation/. Escúchelo utilizando auriculares. Represente su forma de onda utilizando el módulo de Python incluido en el material, appsa_pr1.py.

- ¿Cuál es la frecuencia de muestreo (fs)? 44100
- ¿Cuál es la duración del fichero de audio en segundos? 10
- Incluya la figura obtenida en su memoria. Con motivo de comprensión se ha realizado el proceso con varios ficheros, por lo que se subirán las fotos de ambos

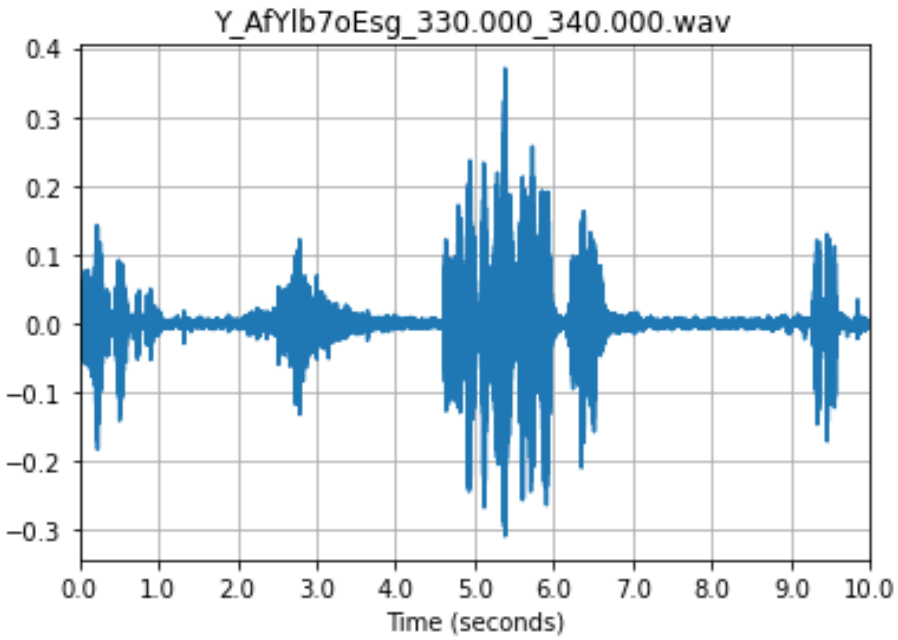

Figura 1.1.- Forma de Onda (Fichero 1 – Agua + Voz)

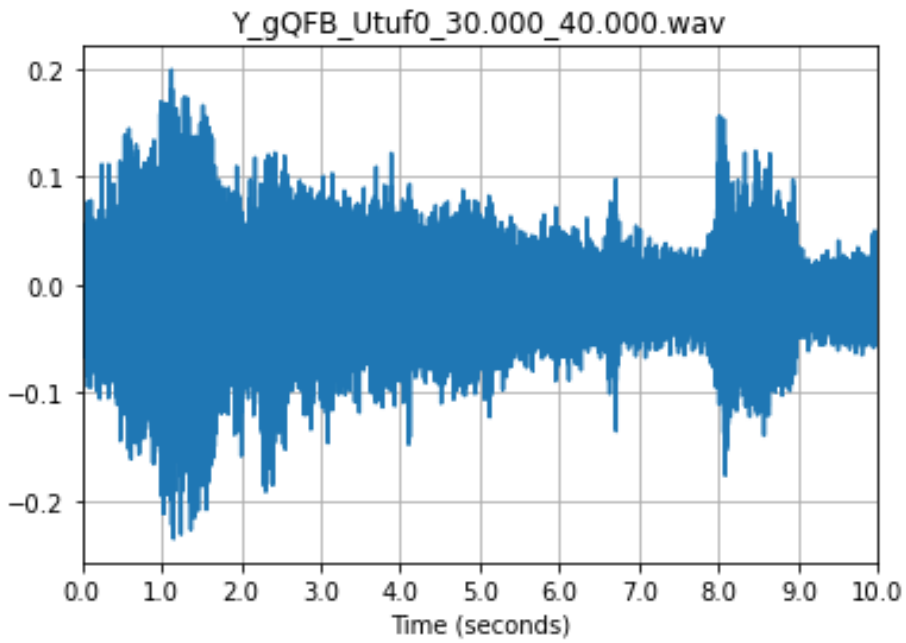

Figura 1.2.- Forma de Onda (Fichero 2 - Gato)

Represente el mel-espectrograma del fichero de audio usando appsa_pr1.py.

- ¿Cuántas ventanas temporales (frames) contiene dicha representación? 864
- ¿Cuántos componentes frecuenciales (filtros mel) se representan en el mel-espectrograma obtenido? 64
- Incluya la figura obtenida en su memoria.

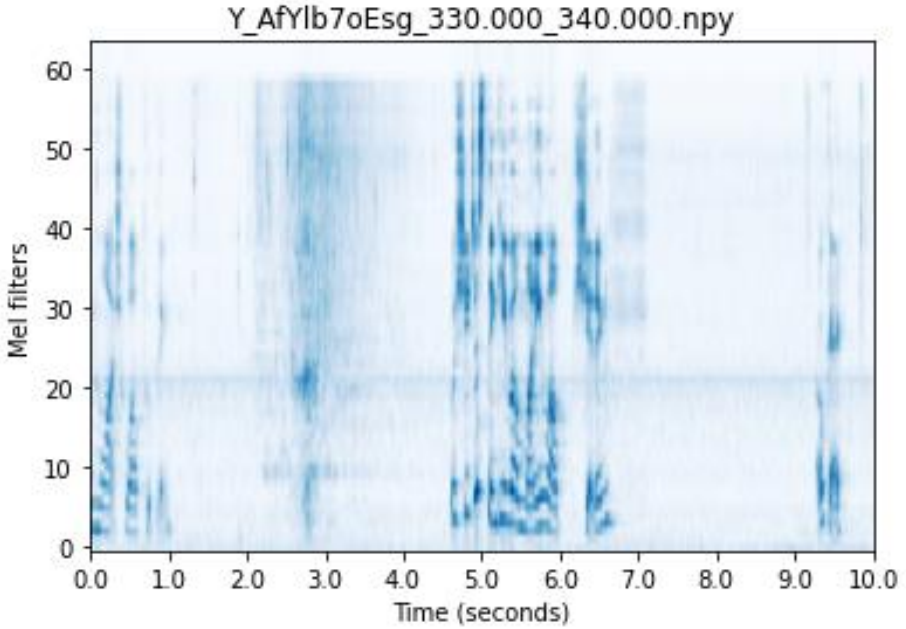

Figura 2.1- Espectrograma de Mel (Fichero 1 – Agua + Voz)

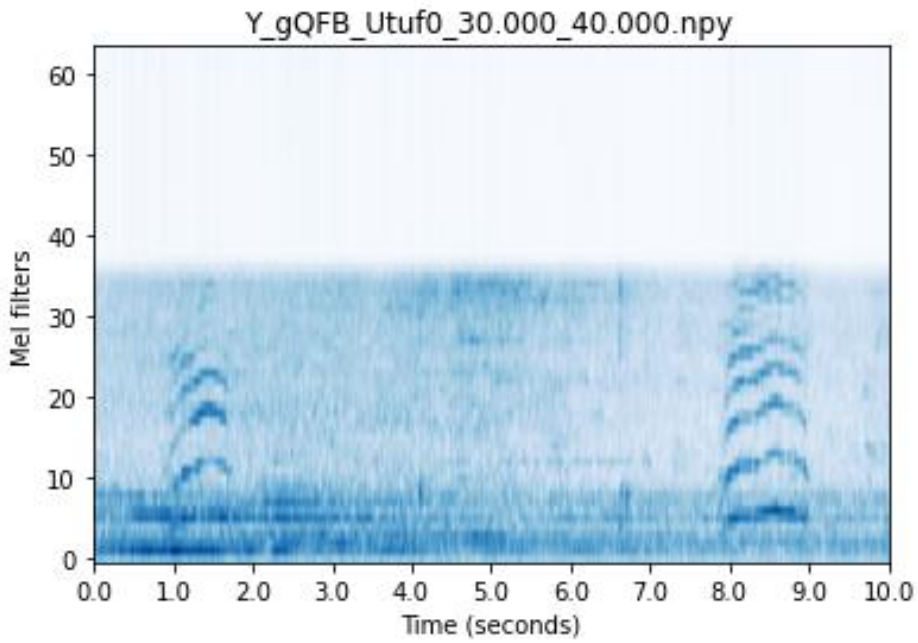

Figura 2.2- Espectrograma de Mel (Fichero 2 - Gato)

- Compare el tamaño de la forma de onda con el del mel-espectrograma, en términos del número de muestras presentes en cada representación.

b) Anotaciones de Eventos

Cargue las anotaciones (etiquetas) de eventos del set de validación (dataset/metadata/validation/validation.tsv) en Spyder o utilizando un editor de textos y busque las etiquetas correspondientes al fichero de audio previamente representado.

- ¿Qué categorías de eventos están presentes en el segmento de audio? ¿En qué instante temporal empieza (onset) y termina (offset) cada una de ellas?

Fichero 1:

Voz/Habla en segmentos de tiempo ([on, off]): [0.0, 1.3], [4.574, 6.676], [9.16, 9.776]

Agua Corriendo: [1.985, 4.574]

Fichero 2:

Gato maullando: [0.894, 1.740999], [7.93, 8.962]

Represente las etiquetas en Python utilizando el material provisto.

- ¿Existe solapamiento de eventos durante el segmento de audio? No
- Incluya la figura obtenida en su memoria.

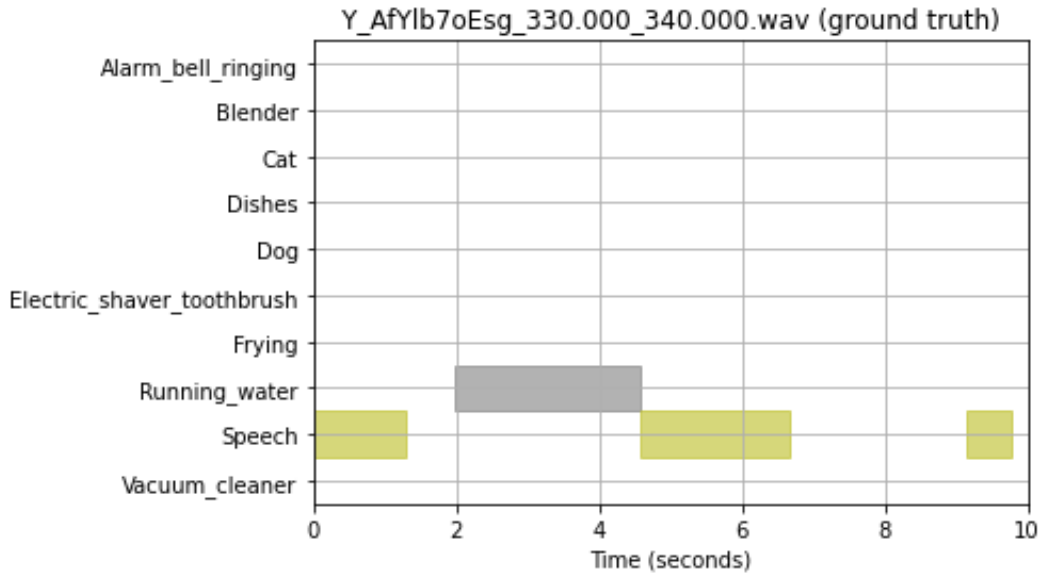

Figura 3.1- Etiquetas del audio (Fichero 1 – Agua + Voz)

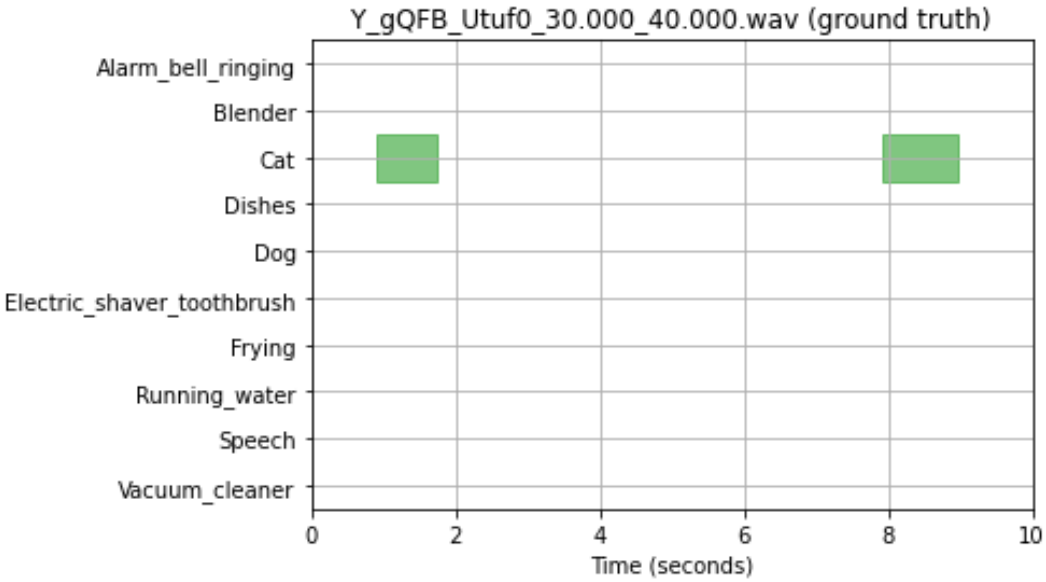

Figura 3.2- Etiquetas del audio (Fichero 2 – Gato)

- Compare la forma de onda y el mel-espectrograma con las anotaciones de eventos, y trate de asociar los eventos de audio anotados con distintas partes de dichas representaciones.

En la Figura 3.1 se ve el etiquetado de eventos en el tiempo. Mirando ahora el espectrograma de mel se pueden reconocer los diferentes patrones que realiza la voz de la persona y el flujo del agua.

En la figura 3.2 se aprecia a la perfección el maullido de un gato durante los intervalos de tiempo: Segundo 1 a Segundo 2, Segundo 8 a Segundo 9. Mirando el espectrograma de Mel de dicho fichero (Figura 2.1), se ve claramente en los mismos intervalos el patrón formado por las frecuencias del maullido

3.- Detección de eventos acústicos con un modelo pre-entrenado.

a) Resultados y métricas

Abra la terminal y ejecute el script de Python TestModel.py para generar predicciones del set de validación con el modelo pre-entrenado.

python TestModel.py --model_path=pretrained_model.p

Se mostrarán los resultados en la línea de comandos, organizados en las siguientes categorías:

- Event based metrics (onset-offset):
  - Overall metrics (micro-average)
  - Class-wise average metrics (macro-average)
  - Class-wise metrics
- Segment based metrics:
  - Overall metrics (micro-average)
  - Class-wise average metrics (macro-average)
  - Class-wise metrics
- Weak F1-score per class
- Weak F1-score macro averaged

Durante la práctica solamente se tendrán en cuenta las secciones resaltadas, que se corresponden con métricas basadas en eventos (event-based).

El F1-score es la métrica principal para evaluación de sistemas de detección de eventos acústicos. Se obtiene como función del número de aciertos positivos (True Positive, TP), falsos negativos (False Negative, FN) y falsos positivos (False Positive, FP) en cada categoría:

$$F1 = (2 \times TP) / (2 \times TP + FP + FN)$$

La puntuación F1 se suele expresar como un porcentaje, de modo que el máximo posible sería 100% (en un caso ideal en el que todos los eventos se detectan de forma correcta sin falsos positivos) y el mínimo sería 0% (cuando ningún evento se detecta de forma correcta).

El rendimiento global del sistema se calcula como el promedio (macro-average) de los F1 de cada categoría.

- Incluya en su memoria los F1 basados en eventos que se obtienen en cada categoría y el F1 promedio (macro-average) obtenido por el modelo pre-entrenado.

Class-wise metrics ==========

F-measure (F1): 25.42%

Event label	Nref	Nsys	F	Pre	Rec	ER	Del	Ins
Running water	237	331	28.9%	24.8%	34.6%	1.70	0.65	1.05
Blender	96	119	23.3%	21.0%	26.0%	1.72	0.74	0.98
Alarm_bell	420	283	38.7%	48.1%	32.4%	1.03	0.68	0.35

Dishes	567	502	13.1%	13.9%	12.3%	1.64	0.88	0.76
Electric _sound	65	100	14.5%	12.0%	18.5%	2.17	0.82	1.35
Dog	570	431	6.6%	7.7%	5.8%	1.64	0.94	0.70
Cat	341	331	20.8%	21.1%	20.5%	1.56	0.79	0.77
Frying	94	376	11.9%	7.4%	29.8%	4.40	0.70	3.70
Vaccum _cleane	92	127	32.0%	27.6%	38.0%	1.62	0.62	1.00
Speech	1754	1213	35.9%	43.9%	30.3%	1.08	0.70	0.39

b) Predicciones

De manera adicional, el script ejecutado en el apartado anterior guardará las predicciones de eventos generadas por el modelo en un fichero TSV con el nombre validation2019_predictions.tsv.

- Utilice el módulo appsa_pr1.py para representar las predicciones del modelo preentrenado correspondientes al segmento de audio elegido en los apartados previos, e incluya la figura obtenida en su memoria.

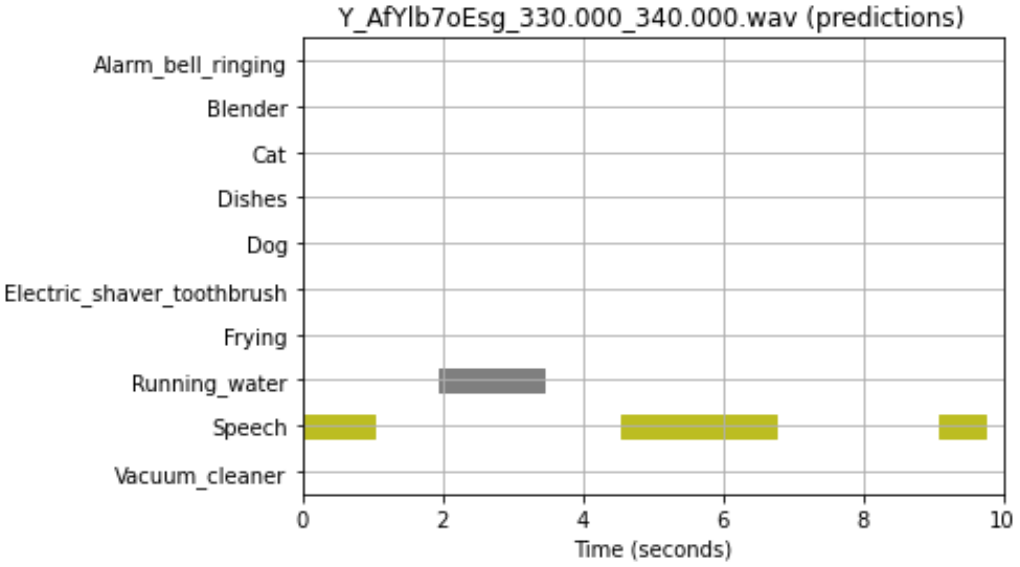

Figura 4.1- Predicción de audio (Fichero 1 – Agua + Voz)

- Comente el ajuste de las predicciones generadas al etiquetado ‘ground truth’.

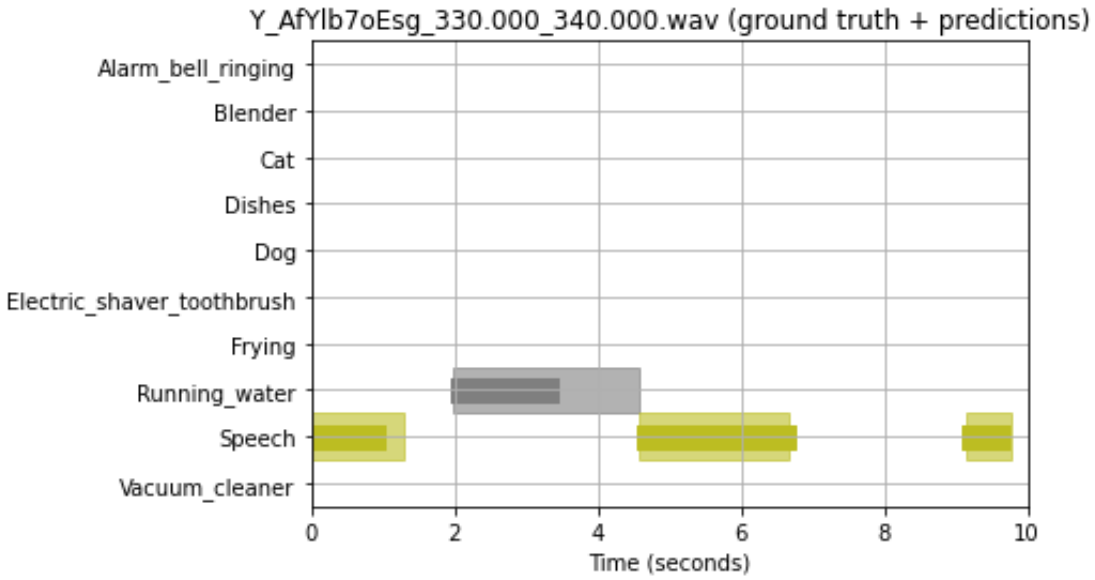

Figura 4.2- Predicción de audio + Etiquetas reales (Fichero 1 – Agua + Voz)

Para este fichero, las predicciones son muy acertadas. El audio en este es muy limpio, se puede diferenciar cada evento en su tiempo, Las predicciones que mas se acercan son las de voz, la que consigue un peor resultado es la predicción del evento de agua corriendo. Esto puede ser debido a es un audio mucho más variable, en cuanto a sonido, incrementa y luego decrementa, en cuanto a el tono o pitch, y en cuanto a otros parámetros que puedan variar.

- Repita el proceso de los dos puntos anteriores para dos segmentos de audio adicionales en los que aparezcan categorías de eventos diferentes.

Fichero 2 – Gato Maullando con ruido de fondo.

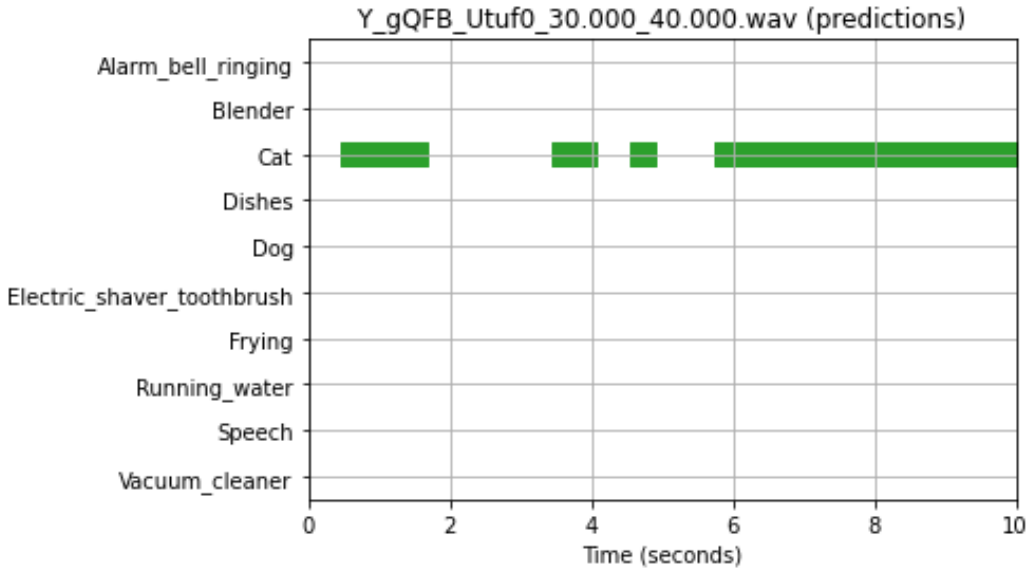

Figura 5.1- Predicción de audio (Fichero 2 – Gato)

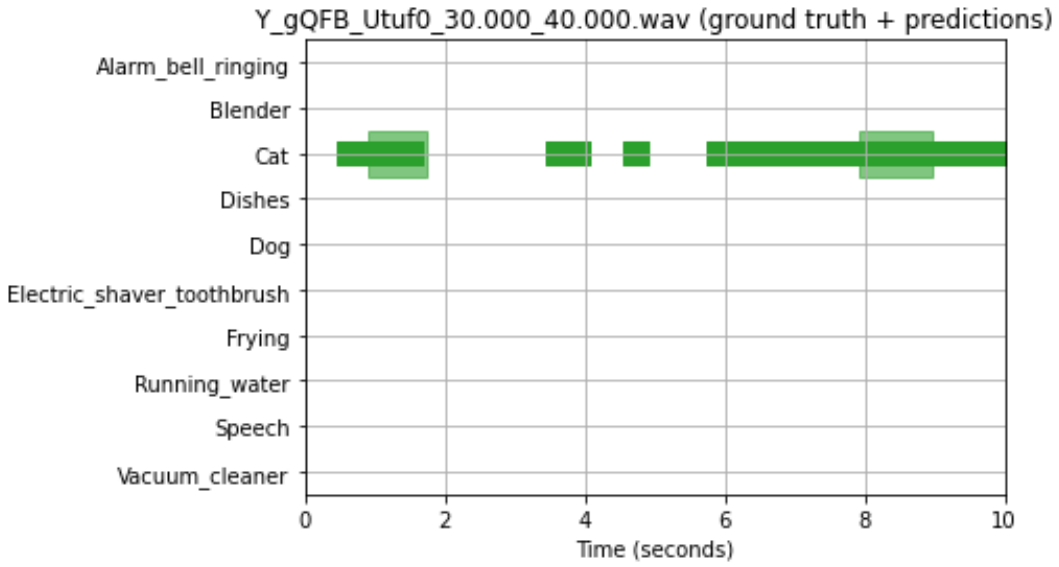

Figura 5.2- Predicción de audio (Fichero 2 – Gato)

En este fichero se observa una predicción bastante mala, en comparación con la del fichero 1. Suponemos que esto es por una entrada más sucia, el audio es grabado con mucho ruido de fondo, y con algún tipo de solapamiento de características entre el maullido del gato y el ruido. Como un pitch elevado o agudo, de manera que el modelo entrenado entienda como maullido un rango de sonidos más amplio y menos especifico.

Fichero 3 – Hombre hablando mientras suena la alarma

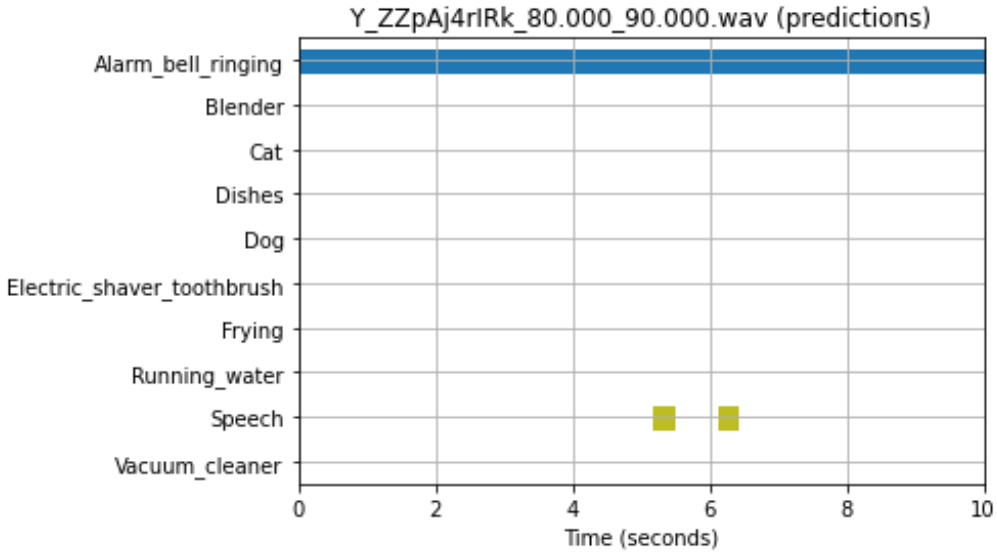

Figura 5.3 - Predicción de audio (Fichero3 – Alarma + Voz)

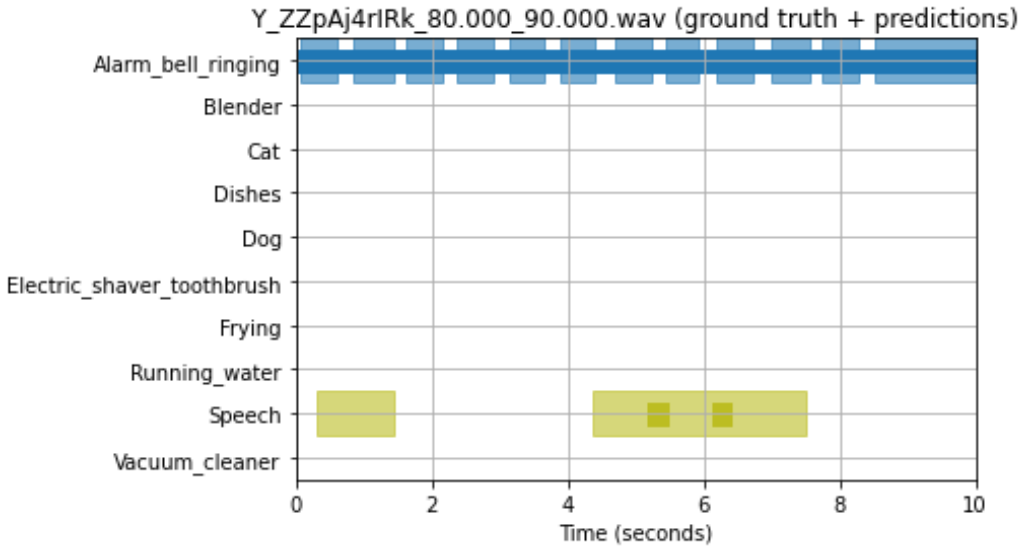

Figura 5.4 - Predicción de audio (Fichero 3 – Alarma + Voz)

En este fichero apreciamos una alarma, repetitiva, cada poco segundo. Observando las predicciones se aprecia que estas no han sido capaces de adaptarse a este evento repetitivo, algo que es más sencillo de predecir, o menos aleatorio, que la cadencia de maullidos de un gato.

De igual modo, al tener la alarma un volumen tan elevado, aunque el etiquetado de voz si es más correcto, las predicciones son bastante más pobres. Casualmente cerca al momento donde se predice voz, en el audio suena más alto, pero esto puede no tener que ver.

En este caso, es en el único estudiado en el que se solapan diferentes eventos de audio. Algo muy importante a tener en cuenta.