Reconocimiento de comandos de voz

Lea completamente esta guía antes de realizar la práctica

Reconocer comandos de voz es una tarea de procesamiento de lenguaje natural (NLP) que implica convertir/transcribir el habla humana en texto. El objetivo es que un sistema informático pueda entender y ejecutar acciones basadas en las instrucciones habladas por un usuario. Este proceso se conoce como reconocimiento automático del habla (ASR, por sus siglas en inglés).
En este caso realizaremos el proceso básico para reconocer comandos aislados de voz que permitan generar una señal que posteriormente pueda ser usada por ejemplo para controlar el estado on/off de la luz. Aquí hay una descripción general de los pasos involucrados en la tarea de reconocimiento de
comandos de voz:

Captura del Audio: El primer paso implica la captura del audio
Deben grabar suficientes ejemplos para cada comando a reconocer. Por ejemplo, queremos reconocer *encender* y *apagar*. Debemos tener suficientes repeticiones de cada caso. Se recomienda más de 30 en cada caso, de ser posible incluso más.
Se adjunta el código para realizar la captura de cada comando en el notebook, seguir las instrucciones al ejecutarlo.

Preprocesamiento del Audio:
El audio capturado generalmente se preprocesa para eliminar ruido no deseado y mejorar la calidad del sonido. Esto puede incluir la cancelación de ruido, normalización y otros pasos para mejorar la claridad de los datos capturados.
No es necesario si se ha realizado en un espacio con bajo nivel de ruido.

Segmentación del Habla:
El audio se divide en segmentos más pequeños que contienen unidades coherentes de habla, como palabras o frases. Este proceso se conoce como segmentación del habla y es esencial para el reconocimiento preciso. En este caso, vamos a eliminar los segmentos de silencio que se puedan presentar en los extremos del archivo grabado. El silencio no aporta información relevante, por lo que se puede eliminar esos segmentos identificando las porciones de la señal que tienen poca energía. Se puede emplear la envolvente de energía o la envolvente de Hilbert.

Extracción de Características:
Se extraen características relevantes del habla en cada segmento, como medidas de energía, entropía, centroide espectral, relaciones de energía entre diferentes bandas…. También se puede usar espectrogramas y otros descriptores que ayudan a representar el contenido del habla de manera adecuada.

En este ejemplo realizamos 2 enfoques: **1**) extraer características puntuales como energía, entropía, centroide espectral…. **2**) usar el espectrograma de Mel y se aplica una técnica de procesamiento adicional para reducir su dimensión.

Modelo de Reconocimiento Automático comandos de voz
Aquí es donde entra en juego el modelo de aprendizaje automático. Utilizando técnicas de aprendizaje automático, como regresión logística o máquinas de soporte vectorial, se entrena un modelo para mapear las características del habla a texto. Durante esta fase de entrenamiento, el modelo aprende a reconocer patrones en los datos de entrenamiento y a asociarlos con transcripciones de habla correspondientes. Es fundamental el uso de un enfoque para el cálculo de características que alimenten el modelo.
Para mantener la simplicidad, usaremos un modelo de regresión logística.

Decodificación:
Una vez entrenado, el modelo se utiliza para decodificar nuevas secuencias de audio y generar transcripciones de texto correspondientes.
Es necesario capturar un nuevo audio realizar todo el proceso y pasarlo por el modelo entrenado en el paso anterior para poder hacer la decodificación

Ejecución de Comandos:
Finalmente, una vez que se ha obtenido la transcripción del habla, el sistema puede ejecutar acciones específicas asociadas con los comandos detectados. Esto podría incluir activar dispositivos, realizar búsquedas en la web, enviar mensajes, entre otros