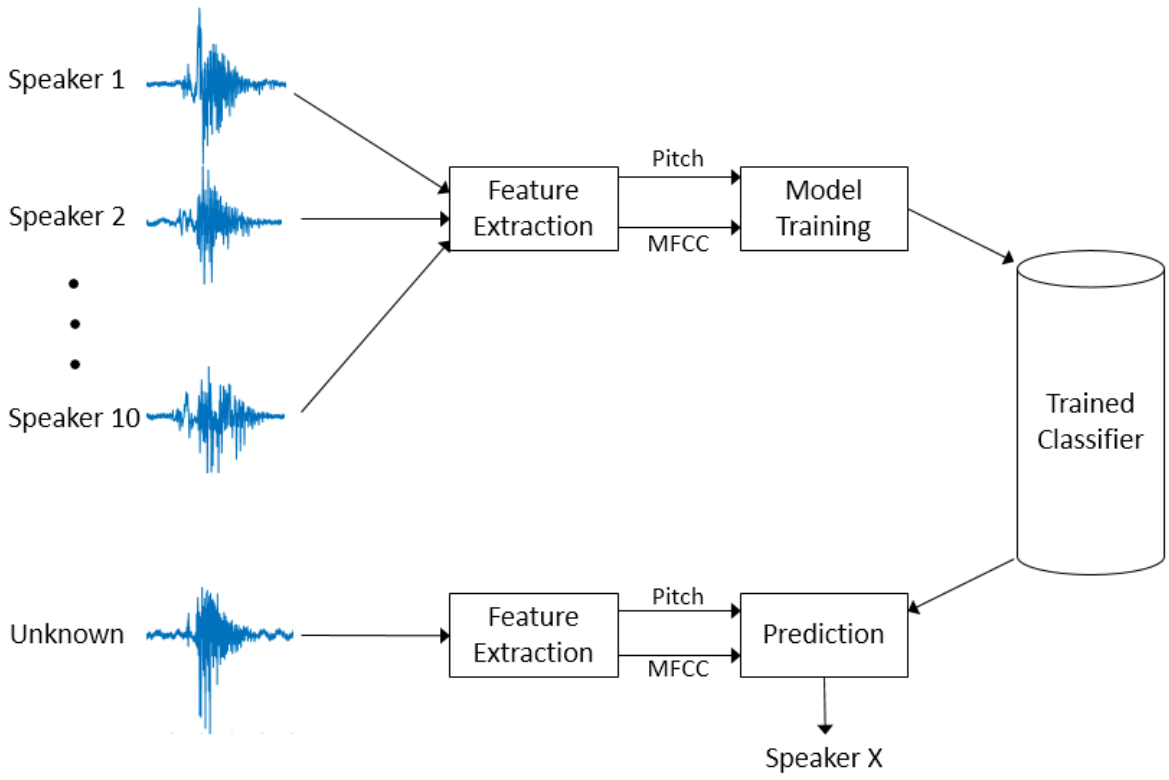
RECONOCIMIENTO DE VOZ

Practica 1 U3

Identificación de hablantes utilizando técnicas temporales Pitch y MFCC

Las características utilizadas para entrenar al clasificador son el tono de los segmentos sonoros del habla y los coeficientes de frecuencia mel cepstrum (MFCC). Se trata de una identificación de hablante de conjunto cerrado: el audio del hablante en prueba se compara con todos los modelos de hablantes disponibles (un conjunto finito) y se devuelve la coincidencia más cercana.

Utilizaremos el siguiente enfoque:



Grabar 10 audios de entre los integrantes del equipo, para generar la base de datos, solo decir números (Solo pronunciar por ejemplo: “uno”, “dos”, “tres”.....hasta el “diez”), de las cuales se extraeran el Pitch y los MFCC, para entrenar un clasificador de K-vecinos más cercanos (KNN). Luego, las nuevas señales de voz que necesitan clasificarse pasan por la misma extracción de características. El clasificador KNN entrenado predice cuál de los 10 hablantes es el más parecido.

RECONOCIMIENTO DE VOZ

Practica 1 U3

La practica consta de dos secciones:

1.- Solo utilizar Pitch y MFCC para entrenar el modelo y para comparar con un audio externo a la base de datos.

2.-Utilizar Pitch y MFCC en conjunto con Zero-crossing rate y short-time energy para determinar cuándo se utiliza la característica de Pitch.

Pitch

El habla se puede clasificar en términos generales como sonora y sorda. En el caso del habla sonora, el aire de los pulmones es modulado por las cuerdas vocales y da como resultado una excitación cuasi periódica. El sonido resultante está dominado por una oscilación de frecuencia relativamente baja, conocida como tono. En el caso del habla sorda, el aire de los pulmones pasa a través de una constricción en el tracto vocal y se convierte en una excitación turbulenta, similar al ruido. En el modelo de fuente-filtro del habla, la excitación se conoce como la fuente y el tracto vocal como el filtro. Caracterizar la fuente es una parte importante de la caracterización del sistema del habla. Como ejemplo de habla sonora y sorda, considere una representación en el dominio del tiempo de la palabra "dos", la consonante /d/ (habla sorda) parece ruido, mientras que la vocal /o/ (habla sonora) se caracteriza por una frecuencia fundamental fuerte.

Una señal de voz es dinámica por naturaleza y cambia con el tiempo. Se supone que las señales de voz son estacionarias en escalas de tiempo cortas y su procesamiento se realiza en ventanas de 20 a 40 ms.

La función de tono estima un valor de tono para cada cuadro. Sin embargo, el tono solo es característico de una fuente en regiones de habla sonora. El método más simple para distinguir entre silencio y habla es analizar la energía de corto plazo. Si la energía en un cuadro está por encima de un umbral determinado, se declara el cuadro como habla.