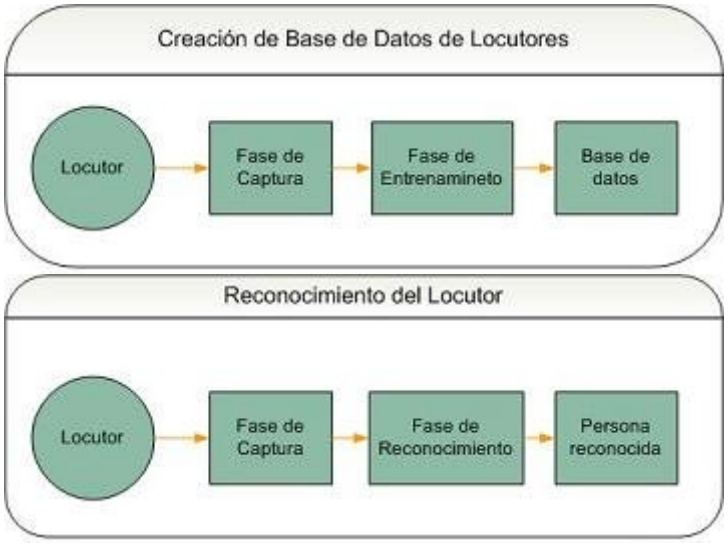
Vamos por partes, como dijo Jack

Las Distintas Fases y Modos de Funcionamiento



Existen dos tareas que realizaremos con nuestro programa:

Modo 1 : Creación de Base de Datos de Locutores

Sin una base de datos, el programa no tendrá a quien reconocer.

Antes de comenzar reconociendo personas, debemos ingresar los datos de uno o más locutores y grabarlos hablando (uno a la vez, lógicamente).

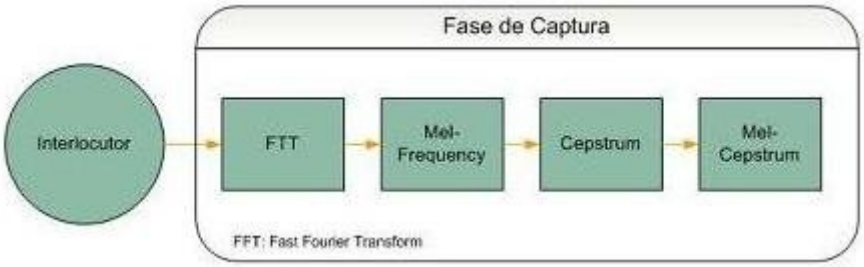
El programa se encargará de reconocer las características de su voz y guardarlas en una base de datos, para futuras comparaciones (reconocimiento).

Modo 2 : Reconocimiento de un Locutor

Una vez que tenemos una base de datos con una o más personas, podremos grabar a un locutor hablando y pedirle al programa que revise si aquella persona está o no en la base de datos. Si la persona se encuentra, entonces el programa dirá de quien se trata.

VoxID te 0ye

Fase de Captura



La entrada a este proceso es un archivo de audio en formato WAV. Este archivo puede haberse encontrado en el disco duro o haber sido grabado en el momento con un micrófono.

El primer paso es dividir esta entrada en ventanas de tiempo. Cada ventana de tiempo tiene un largo definido de puntos de información de audio. Cada ventana será procesada utilizando 3 algoritmos: FFT, cambio de dominio a mel-frequency y cálculo de cepstrum.

Después de que se han aplicado estos 3 algoritmos sobre una ventana, se repite con otra ventana, la cual esta separada a una distancia definida de la otra (suele ser una distancia menor al largo de una ventana).

La ventana no correspode exactamente a la información original contenida en la sección correspondiente del archivo, sino que se le aplica una función hamming para transformarla en una ventana hamming.

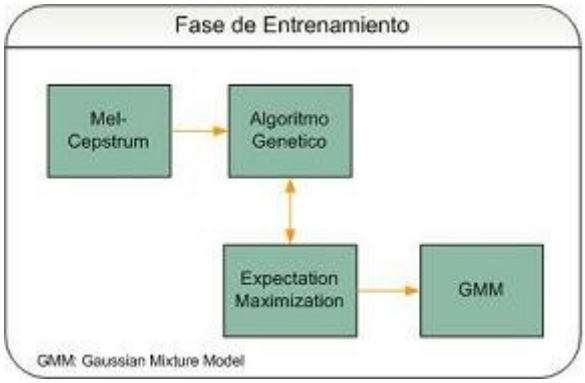
El algoritmo FFT calcula la transformada rápida de Fourier para la señal contenida en la ventana hamming. Los coeficientes de esta transformada se pasan por filtros mel y funciones de frecuencia mel, para cambiar el recorrido de la señal y la ponderación de ciertos coeficientes. Finalmente, se calcula el cepstrum para los coeficientes obtenidos.

El resultado es un vector de coeficientes que puede pensarse como un vector de características de la voz de un individuo.

La transformada de Fourier permite representar las componentes de una señal, lo que podría pensarse como una representación del timbre de voz de una persona. Pasar a dominio de frecuencia mel acerca la señal a la forma como escuchamos los seres humanos. Finalmente, el cálculo del cepstrum permite separar características que son propias de la voz de aquellas que se producen por distorsiones en la cavidad bucal.

VoxID te Modela

Fase de Entrenamiento



El entrenamiento se basa en dos algoritmos de soft computing: algoritmos genéticos y expectation maximization. El segundo es una especie de algoritmo de hill climbing.

El algoritmo genético se basa en la teoría de la evolución de los seres vivos. Primero se inicializa un grupo de GMMs de forma aleatoria. Por cada GMM se calcula su proximidad (fitness) con los datos. Mientras más alto sea el fitness de un GMM, más alta será su probabilidad de cruzarse con otro GMM. El cruce produce una mezcla de gaussianas y sus propiedades entre ambas GMM (también existe la posibilidad de que una GMM pase a la siguiente iteración sin cambio, lo que se conoce como elitismo). Este proceso se produce iterativamete, hasta contar con un GMM con un fitness suficientemente elevado. En cada iteración existe también una posibilidad de que se produzca mutación, en la cual se cambian de forma aleatoria algunas propiedades de un GMM.

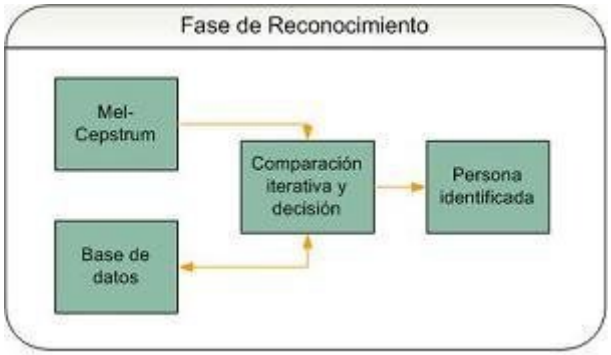
El algoritmo de expectation maximization consiste en una serie de cálculos que permiten obtener un GMM mejor o igual al de la iteración anterior (se basa en derivación y gradiente de funciones).

Esta fase mezcla ambos algoritmos realizando una cierta cantidad de iteraciones de algoritmo genético y después otra cantidad de maximization expectation, lo que se repite varias veces. Esto permite una convergencia a la solución de forma más rápida, al mismo tiempo que, el algoritmo genético, permite evitar estancarse en una solución máxima local.

EL GMM final se almacena en una base de datos.

VoxID te Reconoce

Fase de Reconocimiento



La fase de reconocimiento es bastante simple. Recibe el un set de vectores de características de un locutor desde la fase de captura y se pasan a cada GMM contenido en la base de datos. Donde se repite una operación de cálculo de fitness como la existente en la fase de entrenamiento. Esta operación arroja como resultado un número que corresponde a la cercanía de los datos de entrada con el modelo. Aquél modelo con el número más alto indica quién es la persona más probable, dentro de la base de datos, a corresponder con el locutor. Eso sí, si dicha probabilidad no supera un umbral definido, el programa indica al usuario que el locutor no ha podido ser reconocido. De esta forma, el programa puede indicar también al usuario cuando existe la posibilidad de que el locutor no esté listado en la base de datos.