Capítulo 3. Reconocedores de voz.

Entre todas las aplicaciones descritas anteriormente, llama nuestra atención las que utilizan reconocimiento de voz porque consideramos que son las más útiles a los invidentes, ya que no es necesario aprenderse una serie de comandos para tener acceso y manejo de la computadora.

Por lo anterior, nuestra aplicación se basa en implementar (adaptar) un reconocedor de voz ya existente a un sistema para invidentes (MexVox); de manera que es necesario describir de manera general como funcionan las aplicaciones que utilizan reconocedores de voz.

3.1 El lenguaje hablado.

El lenguaje es una forma convencional utilizada para la comunicación de pensamientos y sentimientos entre los seres humanos. El lenguaje es representado por símbolos ya sean escritos o hablados. En el lenguaje hablado se utiliza la capacidad de articular sonidos generalmente conocidos como la voz.

La voz es parte integral de nuestras vidas. Las personas con capacidades físicas y mentales normales e incluso discapacitados como los invidentes, utilizan la voz como el principal medio de comunicación. A diferencia del lenguaje escrito, el habla puede comunicar necesidades inmediatas.

La voz o también conocida como fonación es el resultado del sonido producido por la salida del aire que, al atravesar las cuerdas vocales de la laringe, las hace vibrar. La voz se define en cuanto a su tono, calidad e intensidad. El tono óptimo y su rango de variación dependen de cada individuo y están determinados por la longitud y masa de las cuerdas vocales. El tono puede alterarse, variando la presión del aire exhalado y la tensión sobre las cuerdas vocales. Esta combinación determina la frecuencia a la que vibran las cuerdas: a mayor frecuencia de vibración, más alto es el tono [8].

Otro aspecto asociado a la voz es la resonancia. Esta se define como la habilidad que tiene una fuente vibrante de sonido para causar que otro objeto vibre. El pecho, garganta, boca y nariz son cámaras de resonancia que amplifican las bandas o frecuencias (número de vibraciones del tono por segundo) formantes contenidas en el sonido generado por las cuerdas vocales. La calidad de la voz depende de la resonancia y de la manera en que vibran las cuerdas vocales, mientras que la intensidad depende de la resonancia y de la fuerza de vibración de las cuerdas [8].

La articulación se refiere a los sonidos del habla que se producen para formar las palabras del lenguaje. La articulación centra su atención en el aparato vocal: garganta (contiene las cuerdas vocales, cuya vibración produce los fonemas), boca-nariz(cavidades de resonancia, refuerzan ciertas frecuencias sonoras), en donde se producen los sonidos del habla.

El habla se articula mediante la interrupción o modelación de los flujos de aire, vocalizados y no vocalizados, a través del movimiento de la lengua, los labios la mandíbula inferior y el paladar. Los dientes se usan para producir algunos sonidos específicos [8].

3.2 Reconocedores de voz

Reconocimiento de voz es el proceso automático de conversión de palabras habladas a palabras escritas [5]. El objetivo del reconocimiento de voz es que las computadoras tengan la capacidad para comprender el lenguaje hablado y una vez entendido puedan ejecutar funciones especificas o almacenar datos. El campo de aplicación de los reconocedores de voz son: la telefonía, sistemas de seguridad, interacción con computadoras, etc.

El reconocimiento de voz generalmente es utilizado como una interfaz entre humano y computadora para algún software. Debe cumplir 3 tareas [8]:

- Preprocesamiento: Convierte la entrada de voz a una forma que el reconocedor pueda procesar, es decir, convertir la señal análoga a digital.
- Reconocimiento: Identifica lo que se dijo (traducción de señal a texto).
- Comunicación: Envía lo reconocido al sistema software de aplicación

Existe una comunicación bilateral en aplicaciones (Ver Figura 3.1), en las que la interfaz de voz está íntimamente relacionada al resto de la aplicación. Estas pueden guiar al reconocedor especificando las palabras o estructuras que el sistema puede utilizar. Otros sistemas sólo tienen una comunicación unilateral.

Figura 3.1 Componentes en una aplicación [8]

Los procesos de pre-procesamiento, reconocimiento y comunicación deberían ser invisibles al usuario de la interfaz. El usuario lo nota de manera indirecta como: certeza en el reconocimiento y velocidad. Estas características las utiliza para evaluar una interfaz de reconocimiento de voz[8].

3.2.1 Clasificación

Los reconocedores se clasifican de varias maneras, dos de estas clasificaciones son:

- De acuerdo a su propósito.

  Los reconocedores de voz se clasifican de acuerdo al fin para el que estén destinados; pueden ser de propósito general (cuando se reconocen palabras de cualquier dominio) y de propósito específico (cuando se reconocen palabras de un dominio en particular) [8].

- De acuerdo al tipo de habla.

  Otro tipo de clasificación es de acuerdo al tipo de habla que reconocen:

  - Reconocedores de habla aislada: Este tipo de reconocedor obliga al locutor a hacer pausas entre las palabras para hacer mas fácil el trabajo del reconocedor, ya que así puede saber el principio y el comienzo de una palabra.
  - Reconocedores de habla continua: En estos reconocedores se trata de emplear la forma más común del habla, la forma continua. Aunque para el reconocedor suele ser más difícil ya que se puede perder entre palabra y palabra.

3.2.2 Ventajas.

Algunas de las ventajas que se tienen al utilizar un reconocedor de voz son: una manera más rápida para la introducción de datos, comodidad al no tener que estar utilizando el teclado o el ratón (mouse), no es necesario tener una parte visual es decir no hay que estar pegados viendo un monitor, poder estar realizando otras actividades manuales o visuales mientras se utiliza la computadora.

3.2.3 Desventajas.

Algunas de las desventajas que se podrían presentar en el reconocimiento de voz pueden ser los problemas de entendimiento en ambientes en donde existe ruido, la pérdida de frecuencias del sonido en el ancho de banda del canal de audio, la fluidez en el habla de las personas ya que esto puede afectar el entendimiento de las palabras y específicamente en el caso de reconocedores de voz de propósito general el vocabulario puede ser inmensamente

extenso, sobre todo en el español y existen muchos modismos lo cual puede crear confusión de palabras.

3.2.4 Datos utilizados para construir reconocedores.

Existen tres tipos de datos utilizados para construir reconocedores:

- Datos del entrenamiento. Se utilizan para construir el reconocedor y ajustar sus parámetros. Dependiendo de la cantidad de información que se tenga, será el resultado del reconocimiento. A mayor información, mayor precisión en el resultado.
- Datos de las pruebas. Se usan para evaluar nuevos algoritmos en la fase de desarrollo del reconocedor.
- Datos de la evaluación. Sirven para medir el funcionamiento del sistema, por lo que esta información debe ser oculta; es decir, que ninguna parte del sistema ha trabajado con este tipo de información. La cantidad de información de pruebas y de evaluación influye en la fiabilidad de los resultados. Por ello, hay que esperar una desviación máxima de los resultados originales cuando se pruebe con un grupo de pruebas distinto [8].

3.2.5 Proceso de reconocimiento de voz.

El proceso de reconocimiento de voz consiste básicamente en transformar una señal a símbolos y darle algún significado al reconocimiento para realizar una acción (Figura 3.2).

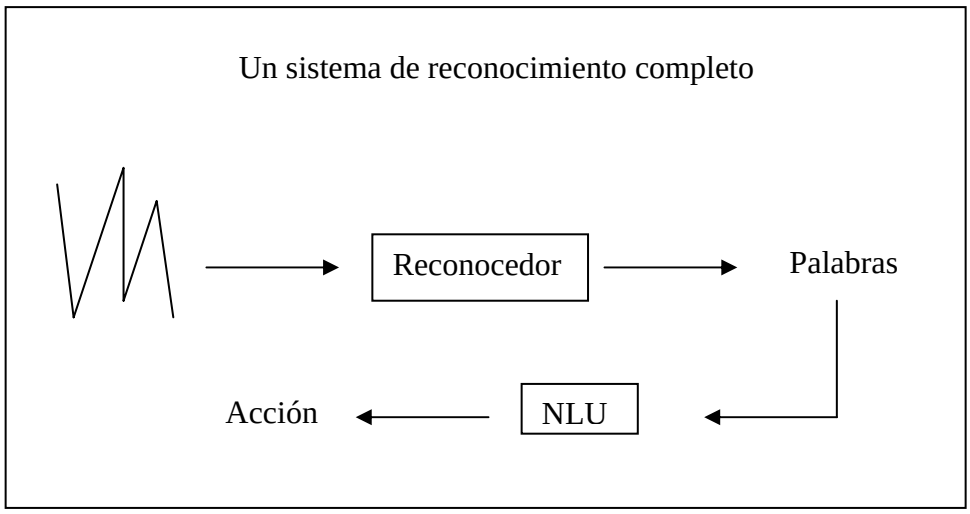

Figura 3.2 Sistema de Reconocimiento de Voz (basado en [8]).

Los pasos para llevar a cabo este reconocimiento son [8]:

1 Obtener los archivos de voz (la señal de voz) y digitalizarlos .

2 Extraer un conjunto de características esenciales de la señal (este conjunto de características será la entrada al clasificador).

3 Introducir el conjunto de características a un clasificador para obtener probabilidades.

4 Búsqueda para encontrar la secuencia permitida más probable. Ya que se tiene las probabilidades y con la ayuda de una estructura que tenga las pronunciaciones posibles, se aplique el algoritmo de búsqueda que dará como resultado el reconocimiento de la palabra.

Estos pasos describen de manera general como funciona un reconocedor de voz independientemente de la tecnología que utilice. Existen varias metodologías para desarrollar reconocedores de voz, dos de las más importantes son las Redes Neuronales

Artificiales y los Modelos Ocultos de Markov las cuales las describiremos de manera general a continuación.

3.3 Redes Neuronales Artificiales.

Las redes neuronales artificiales son modelos matemáticos inspirados en sistemas biológicos que son simulados en computadoras convencionales. Están compuestos de varios nodos simples que operan en paralelo y son arreglados en patrones que simulan redes neuronales biológicas [9].

Las características de las redes neuronales son [9]:

- Habilidad de aprendizaje. (Modifican su comportamiento de acuerdo al medio ambiente).
- Capacidad de generalizar a partir de ejemplos previos.
- Capacidad de abstraer la esencia de una serie de entradas
- Opción de no linealidad.
- Procesan los datos de entrada en paralelo.
- Número y tipo de entradas
- Conectividad de la red.
- Opción de compensación

Ventajas

- Modelos robustos.
- Modelos tolerantes al fallo.

- Pesos de conexión a red no restringidos.
- Implementación rápida (computación en paralelo).

Desventajas

- Requieren la definición de muchos parámetros antes de poder aplicar la metodología mientras que las técnicas estadísticas convencionales, sólo requieren la extracción y normalización de una muestra de datos.

3.4 Modelos Ocultos de Markov.

Los modelos ocultos de Markov son modelos matemáticos basados en probabilidades que pueden ser adaptados para resolver problemas de reconocimiento de voz. Modelo capaz de describir hechos acústicos del habla y que se queda completamente definido por medio de una serie de variables estadísticas [10]. Hay que tener en cuenta una serie de consideraciones previas antes de definir estas variables:

- Un modelo de Markov esta constituido por un cierto número de estados, N, que dependerá del fenómeno que se quiere modelar
- En cada estado, el modelo genera un símbolo perteneciente a un alfabeto finito.
- Las transiciones entre estados pueden producirse cada vez que transcurre un intervalo de tiempo finito igual a la duración de una trama.

Los modelos de Markov están constituidos por dos procesos estocásticos: el oculto, que es el paso de unos estados a otros y el no oculto, que es la generación de símbolos que se produce en cada estado [10].

Ventajas

- Requieren menos memoria física que los de redes neuronales
- Ofrecen un mejor tiempo de respuesta que los de redes neuronales[10].

Desventajas

- Fase de entrenamiento lenta
- Fase de entrenamiento costosa. Pero como esta tarea se realiza una sola vez, vale la pena utilizarlo.