Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

# INTEGRACIÓN DE TAREAS EN AMBIENTE DE RECONOCIMIENTO DE VOZ MULTIMODAL

**Hugo Leonel Zacarías Gómez**

Asesorado por el: Ing. Bayron Wosvely López López

Guatemala, marzo de 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA


FACULTAD DE INGENIERÍA

# INTEGRACIÓN DE TAREAS EN AMBIENTE DE RECONOCIMIENTO DE VOZ MULTIMODAL

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

**HUGO LEONEL ZACARÍAS GÓMEZ**

ASESORADO POR EL: ING. BAYRON WOSVELY LÓPEZ LÓPEZ
AL CONFERÍRSELE EL TÍTULO DE

**INGENIERO EN CIENCIAS Y SISTEMAS**

GUATEMALA, MARZO DE 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

## NÓMINA DE JUNTA DIRECTIVA

| | |
|---|---|
| DECANO | Ing. Murphy Olympo Paiz Recinos |
| VOCAL I | |
| VOCAL II | Lic. Amahán Sánchez Álvarez |
| VOCAL III | Ing. Julio David Galicia Celada |
| VOCAL IV | Br. Kenneth Issur Estrada Ruiz |
| VOCAL V | Br. Elisa Yazminda Vides Leiva |
| SECRETARIA | Inga. Marcia Ivonne Véliz Vargas |

## TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

| | |
|---|---|
| DECANO | Ing. Sydney Alexander Samuels Milson |
| EXAMINADOR | Ing. Otto Amílcar Rodríguez Acosta |
| EXAMINADORA | Inga. Elizabeth Domínguez Alvarado |
| EXAMINADOR | Ing. César Fernández Cáceres |
| SECRETARIO | Ing. Pedro Antonio Aguilar Polanco |

## HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

**INTEGRACIÓN DE TAREAS EN AMBIENTE DE RECONOCIMIENTO DE VOZ MULTIMODAL,**

tema que me fuera asignado por la dirección de la escuela de ingeniería en ciencias y sistemas, con fecha 25 de febrero de 2004.

Hugo Leonel Zacarías Gómez

# DEDICATORIA A:

| | |
|---|---|
| **Mi Dios y mi Madre Celestial:** | por darme la sabiduría y la paciencia para llegar a mi meta. |
| **Mis padres:** | Concepción y Miguel, por brindarme el apoyo económico y moral para llegar a culminar mis estudios. |
| **Mis hermanos:** | Carol y José Miguel, por ser mi soporte y por soportarme en momentos críticos. |
| **Toda mi familia:** | para ustedes este trabajo de graduación. |
| **Mi princesita:** | por impulsarme a culminar mis estudios. |

# AGRADECIMIENTOS A:

**Mi Dios y Mi madre Celestial:** por darme la vida, por saberme guiar por el buen camino, por no abandonarme, por permitirme llegar a este día y alcanzar este objetivo en mi vida.

**Mis padres:** Concepción y Miguel, por ser unos padres ejemplares, por creer en mí y formar valores y principios en todos los aspectos de mi vida y mostrarme con su entereza, dedicación, esfuerzo y amor que son unas personas maravillosas, una bendición de Dios para mí.

**Mis hermanos:** Carol y José Miguel, por entenderme, tenerme paciencia, aguantarme y motivarme siempre a cumplir mis metas.

**Mi familia:** en especial a mis tíos Alis, Bedis, Rosy, Marty, Juan Carlos, Any, y primos Maco, Sergio y mi abuelita Maria (Q.E.P.D.), por todo su apoyo y el ánimo que siempre me inyectaron para lograr esta meta, este logro es de todos ustedes, de toda nuestra familia.

| | |
|---|---|
| **Todos mis amigos:** | en especial a Lester, Marvin, Otto (Q.E.P.D), al QP Team, por todos los desvelos, por escuchar, entenderme, aconsejarme y por ser mis amigos en todo el sentido de la palabra. |
| **Mi asesor:** | Ing. Bayron López, por su apoyo, tiempo y paciencia en revisar este trabajo de graduación. |
| **Mi princesita:** | donde quiera que te encuentres, muchas gracias, todo esto lo he hecho por tí. |
| **La Universidad de San Carlos de Guatemala y sus catedráticos:** | por permitirme adquirir los conocimientos en sus aulas y formar de mí, un profesional íntegro. |

# ÍNDICE GENERAL

# ÍNDICE DE ILUSTRACIONES

## FIGURAS

## TABLAS

# GLOSARIO

**Access®** Gestor de bases de datos desarrollado por Microsoft®.

**Algoritmo** Conjunto de pasos ordenados lógicamente para la resolución de un problema o para la ejecución de algún determinado procedimiento.

**Animación** Dibujos o diagramas en movimiento. Se encuentran frecuentemente en programas didácticos, juegos o presentaciones multimedia.

**API** Acrónico de *Aplication Program Interfase*, que significa interfaz de programa de aplicación. Es un lenguaje formado por un conjunto de instrucciones con determinado formato, utilizado por un programa para comunicarse con otro programa.

**Apple®** Casa fabricante de ordenadores, creadora de ordenadores como los Apple II, Lisa, Macintosh e iMac.

**Árbol binario** Estructura de datos que se representa como un grafo no orientado, compuesto por unidades básicas en las que se distinguen siempre una raíz y dos ramas con hojas.

**Base de datos** Aplicación informática para manejar información en forma de "fichas": clientes, artículos, películas, etc. La mayoría de las bases de datos actuales permiten hacer listados, consultas, crear pantallas de visualización de datos, controlar el acceso de los usuarios, etc.

**Bidimensional** Que tiene dos dimensiones.

**Binario** Significa dos, y es el principio fundamental bajo el cual funcionan las computadoras digitales. Todo lo introducido en el computador es convertido en números binarios formados por los dos dígitos 0 y 1.

**Buffer** Una porción reservada de memoria (de cualquier tipo) que se utiliza para almacenar datos mientras son procesados.

**C** Lenguaje de programación de alto nivel desarrollado en Bell Labs, que es capaz de manipular la computadora a bajo nivel, tal como lo haría un lenguaje Ensamblador. Durante la segunda mitad de la década del 80, el C se convirtió en el lenguaje elegido para el desarrollo del software comercial.

**DOS** Acrónico de *Disk Operating System* (sistema operativo en disco) Sistema operativo monousuario para las computadoras de las series PC, PS/1, PS/2 de IBM.

**Ensamblador** Lenguaje de programación que está a un paso del lenguaje de máquina.

**Frame** El contenido de una pantalla de datos a su espacio equivalente de almacenamiento. Puede definirse también como un cuadro en una animación.

**GLU** Es un agregado que aporta funcionalidad de más alto nivel, que OpenGL no maneja por sí solo. Significa *Graphics Library Unit* o Unidad de librerías gráficas.

**Hardware** Dispositivos electrónicos que almacenan la información electrónica (software), la procesan o ejecutan las instrucciones que ésta especifica. El hardware se puede tocar, el software no, pues no posee sustancia.

**Interactivo(a)** Que tiene un diálogo bilateral entre el usuario y un programa de computadora.

**Interfaz** Una conexión e interacción entre hardware, software y usuario. Las interfaces de hardware son los conectores, cables, etc. que transportan las señales eléctricas en un orden prescrito. Las interfaces de software son los lenguajes, códigos y mensajes que utilizan los programas para comunicarse unos con otros, tal como un programa de aplicación y el sistema operativo.

**Matriz** Colección de elementos en forma de filas y columnas.

**Modelado** Generalmente, es el proceso de representar un objeto real del mundo o fenómeno, como un conjunto de ecuaciones matemáticas. Más específicamente, el término se usa frecuentemente para describir el proceso de representar objetos tridimensionales en una computadora.

**Multiplataforma** Término utilizado, para referir a los programas de software que funcionan en más de un sistema operativo.

**OPENGL** Interfaz para desarrollo de aplicaciones de gráficos creada por Silicon Graphics (SGI), lanzado al mercado por esta empresa en 1992. Partiendo de una interfaz profesional destinada a los más altos niveles en procesamiento gráfico, "una pequeña parte" de esa interfaz ha servido para producir los juegos de video y las maravillas visuales que ahora pueden disfrutarse.

**Puntero** Es una dirección empotrada (incluida) dentro de los datos que especifica la posición de los datos en otro registro o en la memoria. En programación, es una variable que se utiliza como referencia al elemento actual de una tabla, una matriz, arreglo de posiciones (*array*) o algún otro objeto, como la fila actual o columna en la pantalla.

| | |
|---|---|
| **Simulación** | Representación matemática de la interacción de objetos del mundo real. Implmentación de un programa en un lenguaje de máquina, para ejecutarse en una computadora diferente. |
| **SQL** | Leguaje estándar de consulta a bases de datos (*structured query language*, por sus siglas en inglés). |
| **Tridimensional** | Que tiene tres dimensiones. |
| **Vector** | En gráficos de computación, una línea recta designada por sus puntos extremos (coordenadas (x,y) o (x,y,z)). En álgebra matricial, una matriz de una sóla fila o columna. |
| **Windows®** | Sistema operativo creado por Microsoft® que empezó como un entorno operativo que integraba con DOS. |

# RESUMEN

La interfaz es un elemento fundamental en todo sistema de información, ya que de ésta depende la interacción con el usuario y el buen funcionamiento del mismo. En busca de mejorar la comunicación entre el usuario y el computador, se han desarrollado múltiples acciones para que la máquina perciba mejor la información del usuario. Dentro de estas técnicas para "sensibilizar" al computador se encuentra el reconocimiento de voz. Sin embargo, esta técnica no se ajusta totalmente a una conversación entre un computador y un humano, por lo que se ha buscado otro tipo de interacción aunada a la voz, para realizar un reconocimiento correcto y con mayor porcentaje de aceptación.

En este trabajo de graduación se da a conocer el concepto de multimodalidad, aplicado al reconocimiento de voz como opción para potenciar esta técnica, y así mejorar la interacción con los usuarios, independientemente del tipo de usuario con que cuente este tipo de sistemas. Para esto, se desarrolla el concepto de multimodalidad, el cual corresponde a la utilización de varias modalidades para una misma tarea, teniendo como ejemplos prácticos aplicaciones tanto a nivel de informática como fuera de este contexto, ya que es un concepto aplicable en diferentes ámbitos.

Luego, el fusionar este concepto de multimodalidad con la voz, se basa en que el lenguaje hablado es la forma más natural de expresión de los seres humanos. Una persona aprende a hablar antes que a escribir, lo cual hace de la voz, la forma de comunicación más eficiente, ya que incluso se puede utilizar a distancia. Debido a esto, se han desarrollado técnicas de reconocimiento de

voz, entre ellas, el reconocimiento de patrones, redes neuronales y modelos ocultos de Markov.

Contando con todos los elementos para desarrollar el concepto de multimodalidad en ambiente de reconocimiento de voz, se define la metodología como la integración de la voz, con una o más modalidades, para fortalecer el nivel de reconocimiento de la forma más común de comunicación de los seres humanos. Para ésto, se presenta un experimento de una tarea que se trabaja de forma unimodal, es decir, solamente una modalidad, en este caso es manipulación directa frente a esa misma tarea desarrollada con el concepto del reconocimiento de voz multimodal. Esta multimodalidad de voz se logra integrando el reconocimiento de voz, y la manipulación directa para la tarea propuesta.

Definiéndose este experimento y realizándolo con una muestra de 30 personas de diferente sexo, se presentan los resultados obtenidos y las recomendaciones correspondientes para la aplicación de este tema.

# OBJETIVOS

**General**

Diseñar y realizar un experimento de una tarea unimodal, para verificar el nivel de optimización y aceptación del usuario de una interfaz multimodal de reconocimiento de voz.

**Específicos**

1. Evaluación del ingreso de información a una base de datos, como proceso modelo para la verificación de aumento de velocidad, exactitud y aceptación de entradas multimodales cuando las tareas sean separadas o integradas.
2. Estimación de puntos positivos y negativos, de la Implementación y experimentación de una aplicación de ingreso de información, a una base de datos con tecnología de integración de tareas en ambiente de reconocimiento de voz multimodal.
3. Determinación de fortalezas de la modalidad de manipulación directa, para sobreponer las debilidades de la modalidad de reconocimiento de voz y viceversa.
4. Establecer tendencias futuras con relación a la integración de tareas por reconocimiento de voz multimodal.

# INTRODUCCIÓN

La voz, es una forma natural de comunicación que es persuasiva, eficiente y puede ser usada a distancia. Sin embargo, la aceptación amplia de interfaces humano-computadora con voz, es un hecho todavía por ocurrir. Los ambientes multimodales, es un tema que se ha estudiado desde hace algún tiempo. y consiste en presentar un escenario donde se pueda interactuar con diversidad de modalidades, para un proceso o actividad específica. Al tener múltiples tipos de modalidad, éstas se pueden unificar para que un proceso acelere su tiempo de ejecución y pueda llegar a ser muy productivo. Tomando esto en cuenta, varios esfuerzos de investigación se han iniciado para enfocarse sobre la voz, como un canal de entrada auxiliar en ambientes multimodales.

Se denominan tareas multimodales, a las que como su nombre lo indica, se aplican diferentes modalidades de interacción en un ambiente de múltiples modalidades. El objetivo no es simplemente proporcionar dos o más modalidades separadas con la misma funcionalidad, sino integrarlas con la misma funcionalidad para producir una interfaz más productiva. El propósito de ésto, es el de usar las fortalezas de una modalidad para sobreponer las debilidades de la otra. De esta manera, se puede establecer la optimización de un proceso cualquiera, donde las entradas combinadas de lenguaje y del ratón pueden ser más productivas que cualquier modalidad por separado.

Para desarrollar el concepto de reconocimiento de voz multimodal, se presenta en el siguiente trabajo la definición de sistemas multimodales y sus usos actuales, el concepto de la voz y la importancia en la comunicación del ser humano, posteriormente se presenta la tecnología de reconocimiento de voz

pura y sus formas de implementación. A continuación, para demostrar esta metodología, se ha desarrollado un experimento de una tarea para determinar el nivel de optimización de la misma, presentando los resultados obtenidos, con base a un cálculo estadístico de diferencias de tiempos promedios.

El reconocimiento de voz es una técnica por perfeccionar en busca de modelar el comportamiento humano. El estudio de la voz como entrada multimodal, proporciona un cambio al contexto del reconocimiento de voz puro y un aspecto primordial, es el grado de aceptación de los usuarios a una interfaz gráfica multimodal de reconocimiento de voz, aunque aún exista la resistencia al cambio. Bajo todo este argumento, el apoyo en la multimodalidad resulta fundamental, ya que provee funciones que minimizan las deficiencias de la voz y robustecen el nivel de comprensión de una interfaz hombre-máquina.

# 1. SISTEMAS MULTIMODALES

Un sistema multimodal permite la interacción con el usuario a través de diferentes canales de interacción como la voz, gestos, escritura en un teclado o uso de un ratón en una computadora, con el fin de facilitar la obtención del significado de la interacción del usuario.

Debido a que los sistemas multimodales encierran y relacionan varios términos, es necesario definir cada uno de ellos en conceptos. La definición de estos conceptos, se realiza con la finalidad de ampliar el panorama respecto a este tema y con ello, se pueda discutir de forma clara y concisa.

## 1.1 Modalidad

Se refiere al tipo de comunicación usado para obtener información. Se refiere también, a la manera en que una idea es expresada o percibida, o de la manera en que una acción es realizada. Un ejemplo de este concepto, es la percepción de gestos y movimientos en el lenguaje de señas utilizado para personas sordo-mudas. Estos movimientos de las manos y brazos, es la forma de comunicación por la cual, se obtiene y se expresa información.

## 1.2 Modo

Se refiere al estado que determina la forma en que la información es interpretada para extraer o llegar a un significado. Por ejemplo, el modo del lenguaje oral o hablado, es la percepción de los sonidos por medio del órgano

auditivo humano, el cual, determina las palabras y el lenguaje utilizado para brindar el significado debido.

### 1.3 Multimodalidad

Literalmente, “multi” se refiere a “más de uno” y “modal” corresponde tanto a la noción de modalidad como a la noción de modo. El uso de la multimodalidad debe permitir mejorar la comunicación entre varias modalidades de expresión. Esta acción debe cumplir con los siguientes puntos:

- Mejorar el conocimiento en un ambiente ruidoso (audio, video o táctil)
- Establecer una interacción más rápida
- Permitir ser intuitivo y fácil de aprender
- Adaptar diferentes modalidades y diferentes ambientes o diferentes comportamientos de modalidades.

Las modalidades varían de acuerdo a la necesidad de la aplicación, sin embargo, la importancia de la multimodalidad se basa en la existencia de la cooperación entre las modalidades, con el fin, de enriquecer el contenido del mensaje, facilitando la recuperación de la intención del emisor por parte un medio que puede ser por ejemplo, una computadora para ejecutar alguna tarea. El uso de diversas modalidades, implica la fusión de los datos generados a través de cada una de ellas, para lo cual, se necesita tener una técnica de modelado propia de cada una.

En los sistemas de cómputo, el componente que interactúa directamente con el usuario es la interfaz hombre-máquina, y es ésta la responsable de entablar una buena o mala comunicación. Al describir sistemas multimodales

propiamente, se refiere a una interfaz hombre-máquina multimodal ya que son éstas las que interactúan con el usuario.

Entonces se dice, que una interfaz es multimodal, cuando además de utilizar diferentes tipos de datos, es capaz de extraer de ellos el significado de la información. La intención de este tipo de interfaz, es la de aumentar las capacidades de percepción de la computadora. Por ejemplo, cuando dos personas conversan hay más que palabras, también hay gestos y otras formas de expresión. Por lo anterior, el objetivo de las interfaces multimodales es aumentar las capacidades sensoriales de la computadora para mejorar la calidad de la comunicación entre el usuario y el sistema, mediante el uso de todas las modalidades posibles: palabras, gestos, movimientos corporales, etc.

Las interfaces multimodales pueden ser de entrada o de salida, por ejemplo, una interfaz multimodal de entrada es la acción de hablar y apuntar para dar a entender una instrucción específica, y una interfaz multimodal de salida, sería una aplicación multimedia. Las interfaces multimodales tienen componentes generales comunes, los cuales de igual manera, son necesarios aclarar y que se presentan a continuación.

## 1.4 Componentes de interfaces multimodales

Para diseñar y desarrollar interfaces multimodales, es preciso contar con dispositivos o aplicaciones capaces de reconocer las señales producidas de la interacción con el usuario. Estas señales pueden emitirse a través de distintos canales de comunicación, siendo éstos:

- Auditivo
- Visual
- Gestual

Para cada uno de estos canales se debe contar con un dispositivo que reconozca las señales que son emitidas, ya que de ello dependerá el nivel de comunicación que se alcance con el usuario. Mientras se manejen más canales, será más fácil para el usuario interactuar con la máquina.

Actualmente, varios grupos de investigación como Dragon Systems con su software *Naturally Speaking*[1], IBM® con su producto *Via Voice*[2], "Human Interface Technology Labs" con su proyecto HMRS[3] (Sistema de reconocimiento de movimientos manuales, *Hand Motion Gesture Recognition System* por sus siglas en inglés), la "Universidad Carnegie Mellon" en Pittsburgh, Pennsylvania, Estados Unidos, que realiza un proyecto para reconocer patrones que generan voz[4], el "Center for Spoken Language Understanding del Oregon Graduate Institute of Science and Technology" con su *CSLU Toolkit*[5], entre otros, trabajan en diferentes tecnologías para el reconocimiento de estas señales. Algunas de estas tecnologías son reconocimiento de voz, lectura de labios, guantes de datos, apuntadores, marcadores electrónicos, dispositivos sensibles al tacto, etc.

En términos generales, una interfaz hombre-máquina permite establecer un tipo de interacción entre una persona y una computadora. Una interfaz se compone de dispositivos físicos y lógicos que permiten realizar esta tarea. El perfeccionamiento de dichos dispositivos, ha evolucionado de tal forma, que actualmente se pueden encontrar interfaces que manejan entradas y salidas de gráficas, sonidos, lenguaje escrito y hablado, entre otras cosas. Esta forma de

[1] Referencia bibliográfica [16] Dragon naturally speaking

[2] Referencia bibliográfica [36] Via voice

[3] Referencia bibliográfica [20] Hand motion gesture recognition system

[4] Referencia bibliográfica [34] Proyectos multimodales

[5] Referencia bibliográfica [35] CSLU Toolkit

interacción, facilita considerablemente el uso de las computadoras; sin embargo, aún existen aplicaciones de software que no permiten trabajar con este tipo de interfaces multimodales y las herramientas de software que permiten el uso de estas modalidades en las aplicaciones, aún están en desarrollo y son muy limitadas.

Con la finalidad de ampliar el contexto de los campos de acción de las interfaces multimodales, seguidamente se presentan los tipos de componentes que son utilizados para implementar dichas interfaces.

### 1.4.1 Componentes auditivos

El estudio de las interfaces hombre-máquina ha encaminado sus pasos al desarrollo de interfaces que incluyan características propias de las conversaciones humanas, tales como el lenguaje hablado. La variedad de medios y modos de comunicación, enriquece el intercambio de información ya que algunos de estos medios y modos son más eficientes que otros al emplearlos en ciertas tareas y contextos, con ciertos usuarios. Por ejemplo, en una comunicación telefónica predomina el lenguaje.

El campo de las telecomunicaciones ha tenido un gran avance tecnológico, la telefonía digital y los enlaces vía computadora se están convirtiendo en los medios más accesibles y populares de comunicación y de acceso a diferentes servicios. Una forma nueva y natural de controlar algunos servicios es por medio de voz. Ejemplos de algunos de esos servicios controlados por medio de voz son el marcado telefónico, la consulta del saldo bancario vía telefónica, el buzón telefónico de voz y la selección de servicios mediante palabras claves. Estos servicios generalmente incluyen acceso a bases de datos y redes de computadoras.

El uso de este tipo de tecnología con reconocimiento de voz, puede traer otros beneficios como mejorar la productividad de los usuarios de computadoras. Pero de igual forma, es importante hacer referencia que esto puede crear la posibilidad de que la población en general, incluyendo personas con discapacidad física, puedan usar las computadoras, las telecomunicaciones, equipo para el manejo de transacciones, mensajes, información y control de varios dispositivos.

### 1.4.2 Componentes visuales

Los componentes visuales se basan en el reconocimiento de patrones por medio de imágenes. Este tipo de componentes, se utilizan para potenciar otro tipo de modalidades. La razón primordial, es mejorar la tasa de aciertos de interacción de modalidades, para algunos sistemas donde las condiciones no son las óptimas.

Un ejemplo claro de estos componentes es la lectura de labios, como modalidad propia y de igual forma, la utilización de esta modalidad para fortalecer el reconocimiento de voz.

Derivado de lo anterior, se realizan muchas investigaciones en torno a este tema. Por ejemplo, investigadores estadounidenses de la “Universidad Carnegie Mellon” en Pittsburgh, Pennsylvania, desarrollan una computadora que estará en condiciones de leer los labios del usuario y ejecutar sus comandos[6]. La justificación de esta modalidad se basa en el ruido que siempre existe en los entornos o ambientes. Un ejemplo de lo anterior, es un bullicioso

[6] Referencia bibliográfica [34] Proyectos multimodales.

entorno de oficina, donde el sonido de la voz se mezcla con innumerables impulsos auditivos, tanto de otras voces, ruidos provocados por máquinas, etc.

Con base en tal problemática, este grupo de investigadores se interesó al desarrollo de un sistema que hiciera posible interpretar el habla humana con base en el movimiento de los labios y no en el sonido de la voz. La idoneidad del sistema se basa además, en el conocido hecho de que los propios seres humanos, al dialogar en ambientes bulliciosos, también concentran la vista en los labios de su interlocutor. De igual modo, se aspira a perfeccionar la idea del reconocimiento auditivo de voz combinándolo con la lectura de labios. De esta forma, el porcentaje de palabras reconocidas podría acercarse al 100%.

Por otra parte, según científicos de Intel®, en 10 años, los computadores personales podrán entender instrucciones leyendo los labios del usuario con el proyecto AVSR[7] (Reconocimiento del habla audio visual, *Audio-Visual Speech Recognition* por sus siglas en inglés). Ellos desarrollan en la actualidad máquinas que puedan ver y, en consecuencia, leer los labios. Mediante la ayuda de cámaras internas, los computadores de la segunda década de siglo XXI, podrán entender y ejecutar las instrucciones del usuario, siendo necesario que éste sólo se dirija a la máquina y exprese sus deseos. El objetivo de los científicos del centro de investigaciones y desarrollo de Intel®, es que la máquina logre comprender y relacionarse con su entorno visual. Para ello, los participantes en el proyecto desarrollan algoritmos y dispositivos que le permitan a la computadora entender e interactuar con un mundo tridimensional.

Los algoritmos del caso, denominados código estereoscópico, fueron presentados con ejemplos concretos de la tecnología y sus aplicaciones en una

[7] Referencia bibliográfica [10] Visual Interactivity: Audio-Visual Speech Recognition.

biblioteca denominada *OpenCV*[8] (Código abierto de visión para computadoras, *Open Source Computer Vision*, por sus siglas en inglés), la cual fue implementada por Intel® como código abierto para fines de investigación y comerciales. Esta librería, consiste en un conjunto de herramientas de más de 500 funciones de imagen, que ayuda a los desarrolladores a implementar aplicaciones de visión para computadoras. Entre otras cosas, el código puede ser usado para sistemas de seguridad biométricos, en que los rostros, huellas dactilares u ojos de las personas son empleados como códigos de acceso para áreas o sistemas determinados.

En lo que respecta a la lectura de labios por parte de las computadoras, los investigadores consideran que se tratará de la evolución natural de los sistemas actuales, que ya están en condiciones de reconocer la voz humana y seguir sus instrucciones. Así, las computadoras personales del futuro, podrán reconocer el lenguaje humano en una más de sus expresiones.

### 1.4.3 Componentes gestuales

Dentro de los componentes para los sistemas multimodales se encuentran los componentes gestuales. El objetivo de estos componentes, es aumentar la percepción a una interfaz con una computadora por medio de reconocimiento de patrones de gestos o por percepción táctil. Cuando dos personas establecen un diálogo, las modalidades de comunicación son muchas, desde la voz, la lectura de labios, si se encuentran en un ambiente ruidoso y la expresión por gestos. Estos gestos, se pueden realizar con las manos, brazos, con expresiones de la cara, etc. De igual forma, un apoyo para la expresión de ideas en un diálogo, es percibir el entorno por medio del tacto.

[8] Referencia bibliográfica [26] Open Source Computer Vision

Dadas las condiciones anteriores, se realizan investigaciones y productos a nivel de percepción táctil y reconocimiento de gestos para ayudar el funcionamiento de los sistemas multimodales. Un ejemplo de la tecnología respecto a dispositivos sensibles al tacto, es el guante de datos, y dentro de esta categoría, de estos dispositivos actualmente está disponible el *Data-Glove*.

El *Data-Glove*, es el elemento que una persona automáticamente asocia inicialmente al hablar de realidad virtual. Es el dispositivo que permite hacer cualquier cosa en el mundo virtual, así como una mano real puede hacer en el mundo real. El guante de datos es el traductor de lo real a lo virtual. Pero hay demasiada libertad y demasiadas configuraciones posibles, que el usuario simplemente tiene que llegar al punto donde el aparato capte la posición correcta para la ejecución de la acción en el sistema.

Debe realizarse un sistema que deje al usuario la libertad de acción sin la necesidad de una capacitación previa por parte de la interfaz. Por supuesto, algunas acciones son muy intuitivas: agarrar un objeto, apuntar en una dirección, apuntar hacia un objeto. Pero otras órdenes a nivel abstracto, son difíciles de saber por propia intuición. Este problema reduce el poder de los datos, en un aparato en el cual, obtiene la información de un nivel abstracto al ejecutar movimientos digitales.

El estudio de las interfaces hombre-máquina ha encaminado sus pasos al desarrollo de interfaces que incluyan características propias de las conversaciones humanas, tales como el lenguaje hablado. La variedad de medios y modos de comunicación, enriquece el intercambio de información ya que algunos de estos medios y modos son más eficientes que otros al

emplearlos en ciertas tareas y contextos, con ciertos usuarios. Por ejemplo, en una comunicación telefónica predomina el lenguaje.

El campo de las telecomunicaciones ha tenido un gran avance tecnológico, la telefonía digital y los enlaces vía computadora se están convirtiendo en los medios más accesibles y populares de comunicación y de acceso a diferentes servicios. Una forma nueva y natural de controlar algunos servicios es por medio de voz. Ejemplos de algunos de esos servicios controlados por medio de voz son el marcado telefónico, la consulta del saldo bancario vía telefónica, el buzón telefónico de voz y la selección de servicios mediante palabras claves. Estos servicios generalmente incluyen acceso a bases de datos y redes de computadoras.

El uso de este tipo de tecnología con reconocimiento de voz, puede traer otros beneficios como mejorar la productividad de los usuarios de computadoras. Pero de igual forma, es importante hacer referencia que esto puede crear la posibilidad de que la población en general, incluyendo personas con discapacidad física, puedan usar las computadoras, las telecomunicaciones, equipo para el manejo de transacciones, mensajes, información y control de varios dispositivos.

Respecto a la tecnología de reconocimiento de gestos, existen muchas herramientas para la implementación de este tipo de herramientas. Entre estas herramientas se encuentra el *GloveGRASP*[9], que es una librería de lenguaje de programación *C/C++* que permite a los desarrolladores obtener una alta precisión en el reconocimiento de gesto para sus *SGI* (“Silicon Graphics Inc.” por sus siglas en inglés). Estos *SGI*, son dispositivos especiales para

[9] Referencia bibliográfica [18] GloveGrasp

graficación avanzada, el cual provee soluciones tecnológicas de alto rendimiento, visualización y gestión de datos complejos. Este sistema, puede interpretar un gesto de manera diferente para varios contextos de la interacción o un pequeño conjunto de gestos, puede proporcionar un rango grande de órdenes para todos los posibles contextos de la interacción.

Los contextos y los símbolos que representa cada gesto, son completamente definibles por el usuario, permitiendo al reconocimiento dependiente del contexto, reducir la necesidad de recordar docenas de gestos diferentes y produciendo una tasa muy alta de reconocimiento.

Además de estas tecnologías, “Human Interface Technology Labs” está desarrollando traductores que muestran la apariencia gráfica de la entrada emocional dada por el usuario desde el teclado, un ejemplo de esto es *Intelligent Conversational Avatar*[10], el cual es un sistema experto. El parser del sistema experto *Intelligent Conversational Avatar*, toma el texto de las emociones que el usuario desea retratar y aloja señales de cuenta en texto como: los tipos de palabras que se usaron, la información contextual, la longitud de frases que tecleó, uso de emociones o produciendo una salida gráfica en un avatar.

Avatar a nivel esotérico, significa un ser espiritual que “desciende” en respuesta a la llamada y necesidad de la humanidad. Análogamente en el mundo de la informática, un avatar es un facsímile gráfico que se puede utilizar en las habitaciones de discusión virtual. Se puede definir como un personaje tridimensional que puede representarse en los mundos virtuales, la encarnación de un ser humano en un mundo virtual.

[10] Referencia bibliográfica [21] Intelligent Conversational Avatar

*Intelligent Conversational Avatar* es más que un traductor, puesto que este sistema puede producir conversación emocional, ya que da una retroalimentación tanto gestual como visual.

Hasta ahora, se ha expuesto a las interfaces multimodales en concepto y con los componentes que pueden constituirla. El concepto multimodal es amplio y debido a esto, es preciso ejemplificar algunos de los usos de sistemas multimodales, abarcando todo el concepto que éste conlleva.

## 1.5 Uso de sistemas multimodales

La utilización de los sistemas multimodales, puede ser adaptable según el contexto en el que se trabaje. El concepto de multimodalidad, se puede emplear en varios aspectos de la vida, no sólo a nivel de una interfaz humano – computadora. Por ejemplo, se puede observar en una situación tan simple entre 2 personas, en la cual, una le debe indicar a la otra cómo llegar a un lugar en la ciudad. En este escenario, no se describe cada paso, sólo se le dice en qué esquina hay que dar vuelta o qué autobús tomar. Para éste caso, se utiliza se utiliza la multimodalidad en hablar y señalar para dar a entender el mensaje.

Así como ésta situación tan sencilla y común que se presenta en cualquier momento y en cualquier lugar, existen otros entornos donde los sistemas multimodales se hacen presentes. A continuación, se muestran algunos ejemplos de los usos de estos sistemas, tanto para la utilización de una interfaz humano – computadora, como para medios donde no se involucre este tipo de tecnologia y en varios campos de acción como medicina, logística de transportación, mapas interactivos, simulación, sistemas de seguridad y navegación por internet.

### 1.5.1 Sistema de transporte de mercancías

En un mundo con una economía globalizada, el mercado requiere de un transporte moderno y competitivo, que garantice la participación de los productos en condiciones de competitividad. Actualmente, la logística de transporte se centra en la fortaleza del conjunto de conocimientos del mercado, de las operaciones de carga, de los trámites documentales y de la asesoría a los clientes, ya que el éxito del transportador moderno depende del éxito de sus clientes. Para lograr de una forma efectiva la mezcla de estos conocimientos, se hace necesario innovar a sistemas de transporte multimodales de transporte.

Un sistema de transporte multimodal, se basan en el envío de mercancías utilizando varias modalidades de transporte, bajo un documento de envío combinado origen-destino, con un régimen fraccionado de responsabilidad, pues a cada modalidad de transporte se le aplican las normas que lo regulen.

La coordinación del servicio de transporte, la realiza directamente el generador de la carga, o a través de su representante, quien contrata, en nombre del generador, al transportador carretero, ferroviario, aéreo, marítimo o fluvial y realizando el transporte en cualquiera de las combinaciones que mejor se ajusten. Esta forma de contratar el servicio de transferencia, se utiliza habitualmente en el transporte internacional de mercancías y más frecuentemente cuando el trayecto es marítimo.

Debido a lo expuesto anteriormente, los puertos deben de seguir la misma logística, mantener una multimodalidad en sus operaciones. Un ejemplo claro,

es el puerto de Barcelona, España[11]. La idea que tiene la mayoría de la gente, es que cuando piensa en la palabra “puerto” es la de muchos barcos cargando y descargando cosas, pero hay mucho más que esto.

El puerto de Barcelona es responsable de la carga y descarga de productos de vía marítima, aérea, terrestre y ferroviaria. El transporte de entrada o de salida de estos productos se realiza en cualquiera de estas modalidades o en la combinación que se desee. Esto ha logrado elevar el nivel de este puerto a una logística multimodal, la cual trae beneficios al comercio de esta región de España.

Un sistema de transporte multimodal, acompañado de los servicios que faciliten su operatividad, es un factor importante con el cual, es posible mejorar los niveles de competitividad de los productos en los mercados internacionales, ofreciendo una ventaja con la fluidez y rapidez de los envios.

### 1.5.2 Sistemas de análisis psicológicos

Siguiendo con la ilustración de sistemas multimodales, vale la pena hacer mención de la aplicación de éste concepto en la medicina y específicamente en la psicología. Esta idea se maneja como “Terapia Multimodal”[12]. Se plantea la terapia multimodal como un enfoque de conocimiento-conductual alternativo a otros enfoques conductuales y a las orientaciones basadas en la descripción, diferenciación y clasificación de enfermedades que se utiliza.

Se priorizan las ventajas de este enfoque sobre aquellos tanto en la vertiente de evaluación como de tratamiento. Igualmente se hace mención al

[11] Referencia bibliográfica [28] Puerto de Barcelona, España

[12] Referencia bibliográfica [23] Evaluación y terapia multimodal

alza de los enfoques multimodales sobre los unimodales en el campo de la salud mental. El enfoque multimodal de evaluación, tiene cada vez más aceptación en campos como la psicología clínica, la psiquiatría y la enfermería psiquiátrica.

El sistema multimodal trata de evaluar la "personalidad total" del paciente cubriendo su perfil C.A.S.I.Co (C: Cognición, A: Afecto, S: base Somática, I: Interpersonal y Co: Conducta). Parte de la concepción de que los seres humanos tienen imágenes y pensamientos (cognición), afectos, sensaciones sobre una base somática y conductas en contextos interpersonales. Cada caso puede ser evaluado en las anteriores dimensiones y las interacciones que se producen entre ellas, dando con ello, una verdadera importancia a la multimodalidad.

### 1.5.3 Sistemas de análisis de imágenes médicas

Una herramienta fundamental para el diagnóstico de una enfermedad o la planificación de una intervención quirúrgica, son las imágenes médicas digitales. La información que proveen estas imágenes, a menudo, son vitales para extraer datos cuantitativos que permiten saber el grado exacto de afectación de alguna dolencia, e incluso ofrecen pautas para una aproximación del tratamiento correspondiente. Sin embargo, en algunos casos, esta información puede ser insuficiente o poco objetiva. Debido a esto, grupos de investigación trabajan en tecnologías que permitan mejorar la calidad de imágenes, pero por encima de todo, que éstas proporcionen ya mayor cantidad de información posible.

Uno de estos de grupos de investigación se encuentra en el hospital “Gregorio Marañón”, en el laboratorio de imagen médica, de la unidad de

medicina experimental[13]. Este grupo, busca generar plataformas en las que se integren datos, que en la actualidad se obtienen por separado y que faciliten el alcance de un mejor nivel cualitativo para los diagnósticos. Para este efecto, se utiliza la imagen médica “multimodal”. En esencia, consiste en integrar la información obtenida por dos o más técnicas de imagen en una sola.

Es de especial interés la combinación entre técnicas que ofrecen información morfológica, como la tomografía axial computerizada (TAC) o la resonancia magnética nuclear (RMN) y técnicas funcionales, como la tomografía por emisión de positrones (PET). Las primeras dos técnicas permiten localizar estructuras con un alto nivel de precisión geométrica, mientras que la segunda ofrece información sobre el metabolismo de áreas concretas del organismo, aunque con poco detalle morfológico. Esto ofrece una interpretación cuantitativa, además de cualitativa, a las variaciones de gris o colores que aparecen en cualquier imagen.

De igual forma, el valor de la integración, va más allá del diagnóstico. La fusión multimodalidad demuestra su utilidad en terapia. Un ejemplo de lo anterior, es la radioterapia, ya que facilita el cálculo de dosis en las áreas objetivo y zonas sensibles circundantes a proteger. Esta tecnología también puede ser válida para planificación quirúrgica, modelando el área a operar previamente a la intervención.

La tecnología multimodal también está demostrando su validez para la preparación de intervenciones en neurocirugía. Este podría ser el caso de la extirpación de un tumor cerebral, intervención en la que la "cirugía virtual" permite precisar con antelación el riesgo de una eventual extirpación y prever

[13] Referencia bibliográfica [17] Fusión de imagen multimodalidad

zonas colindantes afectadas. El mismo principio se está aplicando con eficacia en el estudio de la epilepsia temporal, una enfermedad en la que se observan cambios de actividad en determinadas zonas del cerebro cuya visualización precisa es mucho mayor gracias a la integración de distintas técnicas de imagen.

En general, las fórmulas que maneja este grupo de investigación tratan de aunar conocimientos de al menos tres áreas: imagen médica digital, bioquímica e informática, aplicando como ventaja, la multimodalidad en la fusión de imágenes para diagnósticos y terapias a seguir.

### 1.5.4 Sistemas de seguridad biométricos

En la actualidad, el acceso seguro a los sistemas es un tema cada vez más importante. En este sentido, se hacen grandes esfuerzos y se gastan enormes sumas de dinero en aspectos de seguridad. Sin embargo, los usuarios no autorizados siguen accediendo a información delicada y colocando a las empresas en posiciones vulnerables.

En la búsqueda de parámetros más confiables, naturales y fiables, se encuentran los parámetros biométricos. La biométrica es la tecnología que mide las características biológicas que permiten la identificación de individuos. Los importantes avances conseguidos en el desarrollo de las tecnologías de reconocimiento biométrico impulsan su incorporación en aplicaciones sobre escenarios reales. Entre los parámetros biométricos utilizados actualmente, podemos destacar las huellas dactilares, la forma del iris ocular, la voz y reconocimiento facial.

No obstante, existen dificultades para la obtención de datos por intermedio de estas técnicas, ya que una sola medición no funciona bien para

todo el mundo. Por ejemplo, la medición de las huellas digitales pueden ser ilegibles o el reconocimiento facial puede fallar si se hacen cirugías, cambios de peinado o de iluminación.

Estas dificultades, propician la utilización de métodos para vulnerar la seguridad de los sistemas aprovechándose de estas debilidades, por ejemplo, la recolección de una huella dejada en un lector para usarla en contra del individuo que la dejó. Por ello, se ha creado la combinación de varias técnicas biométricas, en lo que se ha denominado biométrica multimodal. Este método asigna una probabilidad a cada medición y las combina para obtener un resultado positivo o negativo.

De cualquier forma, existen empresas que han puesto en marcha proyectos utilizando la biometría multimodal. Prueba de ello, es la empresa alemana "HumanScan" que provee el producto *BioId*[14] que se basa en el reconocimiento facial, reconocimiento de voz y movimiento de los labios para identificar a las personas. *BioID*, es una solución avanzada orientada al reconocimiento de personas, con la cual, no se necesita ingresar contraseña, ni ninguna clase de *PIN*.

El *BioID* provee una verdadera multimodalidad, basándose en la habilidad de análisis simultáneo de 3 áreas biométricas: reconocimiento de rostro, voz y movimiento de labios, con lo cual, el usuario simplemente se pone enfrente de una cámara de video y con decir el nombre esta persona puede autenticarse. Este sistema requiere únicamente de una cámara de video y un micrófono y es aplicable para autenticación para accesos físicos a lugares

[14] Referencia bibliográfica [13] BioId

específicos, entradas a redes, consulta de datos personales como correo electrónico y transacciones comerciales.

La seguridad biométrica aun tiene campo por recorrer antes de que se pueda popularizar y masificar, pero se hacen esfuerzos para implementar soluciones con ésta técnica, prueba de ello es el *BioID*. Para la difusión adecuada de la biométrica multimodal, es importante que se definan metodologías de diseño, estándares de desarrollo y aunar esfuerzo en el estudio e investigación a evaluar el impacto real de dichas tecnologías sobre amplias poblaciones de usuarios en escenarios reales.

### 1.5.5 Sistemas de simulación

En el mundo actual, tanto en el área de los negocios, como en la industria y los gobiernos, los proyectos en gran escala y de gran complejidad son la regla y no la excepción. Estos proyectos complejos requieren estudios previos a su construcción o modificación, denominados estudios pilotos.

Para tales estudios pilotos, se realiza la construcción de modelos donde se realiza el estudio con el fin de obtener conclusiones aplicables al sistema real. Construido el modelo, se montan un conjunto de alternativas que se definen para su ensayo, lo que constituye la estrategia de la simulación.

La simulación tiene como principal objetivo la predicción, es decir, puede mostrar lo que sucederá en un sistema real cuando se realicen determinados cambios bajo determinadas condiciones.

Con los avances tecnológicos en el área informática, se ha dado un gran auge en la utilización de la simulación como auxiliar importante en la concreción

de proyectos. La realización de programas que representen sistemas a estudiar y ensayar alternativas, no es otra cosa que dar los datos a una computadora para que ésta imprima los resultados.

Sobre esta línea informática, se han llegado a construir modelos de simulación donde la multimodalidad se hace presente. Como muestra de ello, se presenta la interfaz multimodal de simulacion de sistemas militares llamada *QuickSet*[15]. Esta es una aplicación militar que combina el reconocimiento de voz y la entrada del gesto basado en la pluma electrónica, y permite al usuario preparar ejercicios de entrenamiento creando fuerzas y medidas de mando, y controlar el ejercicio asignando tareas a las fuerzas en *LetherNet*, que es un sistema para simulación desarrollado para entrenamiento.

También le permite al usuario controlar el punto de vista de la unidad de visualización del terreno, junto con sus muchos rasgos (radar, HUD por sus siglas en ingles *Head Up Display*, es un dispositivo de los aviones que brinda los datos básicos y tácticos de vuelo en forma holográfica., etc.), y también proporciona mando de video que cambia para los despliegues de pantalla grande. El sistema consiste en un asistente digital portátil (PDA por sus siglas en inglés) que es un computador personal portátil que pesa 3 libras aproximadamente, empleando aviones comunicaciones **LAN** (área de red local, *Local Area Network*, por sus siglas en inglés) inalámbricas, pantalla a color, micrófono, aguja de la pluma, reconocimiento de voz y reconocimiento de gesto.

La entrada de voz y la entrada de la pluma, tienen cada una tiene sus ventajas y desventajas. La voz habilita nombrando cosas que actualmente no

[15] Referencia bibliográfica [29] QuickSet

estén visibles en la pantalla, como pelotones, o tareas, procedimientos, reglas, o situaciones. La voz también es más rápida para los órdenes emisores. La entrada de la pluma es a menudo más conveniente (y más exacto) para indicar objetos que están actualmente a la vista en la pantalla. La entrada de la pluma también permite especificación de líneas irregulares que podrían indicar rutas, o límites de áreas como campos de minas, pantanos, zonas de aterrizaje, áreas de ensamblaje, etc.

La entrada de la pluma tiene la ventaja de ser utilizable en lugares públicos, donde al usuario no le gustaría poner en palabras órdenes debido a lo privado o secreto de éstas, y también es utilizable donde el ruido de armas, aviones, y vehículos de tierra puede interferir en el uso de un reconocedor de voz. Este sistema de entrada multimodal le da la habilidad de capitalizar ambos juegos de ventajas al usuario y usa cualquier modalidad que satisface la necesidad del momento.

Una interfaz multimodal inteligente emula la comunicación multimodal humana en las interacciones hombre-máquina. Los desarrolladores de estas interfaces aplican herramientas de inteligencia srtificial al desarrollo de la tecnología de interfaces hombre-máquina para integrar lenguaje natural escrito y hablado, gráficas y señalamientos en un diálogo interactivo entre el usuario y la computadora.

### 1.5.6 Sistemas de navegación por internet

El avance de la informática, de las telecomunicaciones y en concreto de Internet, ha hecho posible que un bien tan preciado como la información, esté disponible desde cualquier lugar del mundo permanentemente. Desde su creación, Internet ha ofrecido una amplia gama de posibilidades para el acceso

a esta información. Sin embargo, que la información esté disponible no significa que sea accesible para todas las personas que la pretenden, al menos, no en el mismo grado.

Un ejemplo de lo anterior son las personas con discapacidad, ya que a menudo encuentran dificultades cuando intentan recuperar información de la red. Algunos diseñadores e interfaces han reemplazado la funcionalidad y sencillez por la estética, dificultando el acceso a los contenidos, especialmente para aquellas personas cuyas discapacidades físicas les impide disfrutar con el diseño, como lo son los invidentes.

En cualquier área de la actividad humana, se puede percibir que en un gran porcentaje de la información llega a través de la vista, pero para una persona con ceguera total, el oído y el tacto pasan a ser los principales canales en la recepción de la información, mientras que para las personas con deficiencia visual el resto de visión que poseen es un recurso más a utilizar. Debido a esto, se han desarrollado herramientas para que las personas invidentes, puedan percibir el mundo y su información a partir de otros canales de recepción que no sea la vista. Estas herramientas se basan en fortalecer los medios de comunicación táctiles y auditivos.

El grupo “Quercus de ingeniería del software” de la Universidad de Extremadura, España ha desarrollado proyectos de investigación, en diferentes líneas de investigación dentro del ámbito de la Ingeniería de software. Dentro de los proyectos realizados por este grupo de investigación, se encuentra una herramienta llamada KAI (Kit de Accesibilidad a Internet).

KAI es una herramienta integrada por componentes software y hardware. Su principal meta es la de incrementar la accesibilidad a los contenidos de

internet para usuarios con deficiencias visuales, especialmente los invidentes. La arquitectura de KAI tiene dos componentes principales: el lenguaje *BML* y el navegador *WebTouch*.

El *BML*, (*Blind Markup Lenguage* por sus siglas en inglés) ha sido desarrollado para construir fácilmente páginas de internet accesibles. Se deriva de *XML* (*Extensible Markup Language* por sus siglas en inglés) y es bastante similar a *HTML* (*Hiper Text Markup Language* por sus siglas en inglés) pero incluyendo nuevas etiquetas para mejorar la estructura de la página. *BML* también permite la posibilidad de asignar maneras diferentes de presentación de los diferentes elementos de la página de internet: textual, gráfico, sonido y táctil. El KAI permite traducir una página de internet existente escrita en *HTML* o *XHMTL* (fusión de XML y HTML) a BML, esto, con el fin de que el *WebTouch* pueda interpretarla.

El *WebTouch*, es un navegador multimodal especialmente desarrollado para usuarios invidentes. Proporciona varias posibilidades de navegación combinando audio, voz y habilidades táctiles. Esto permite configurar el modo de navegación de acuerdo a las necesidades del usuario. *WebTouch* permite obtener la representación mediante iconos “táctiles” de los contenidos de la página web en el área de representación. Para esta representación se realiza una traducción de cada página web convencional a otra equivalente pero formada por iconos.

Una vez que el icono ha sido seleccionado, entonces el usuario puede interactuar con él usando un menú contextual. Por ejemplo, dado un texto, el usuario puede activar el control para leerlo. Para ello *WebTouch* dispone de un sintetizador de voz que además ofrece al usuario cualquier información requerida en un determinado momento. El usuario puede también usar el área

de Navegación para navegar por la página a través de diferentes estructuras navegando sólo por un tipo de elementos: sólo tablas, sólo direcciones de e-mail y así sucesivamente.

Uno de los principales objetivos en el diseño de KAI es la facilidad de uso para los usuarios. El usuario sólo necesita un ratón especial y el software desarrollado. Según este grupo de investigación, los experimentos desarrollados hasta ahora han demostrado que la herramienta es muy intuitiva para los usuarios.

Estos han sido algunos ejemplos, de la utilización actual de sistemas multimodales y como se puede observar, el concepto de multimodalidad se presenta en muchos casos de la vida diaria, brindando robustez y facilidad a los campos donde se aplica. Bajo este término de beneficio en estas áreas de aplicación, es necesario extender la perspectiva y detallar tanto las ventajas, como las desventajas que ofrecen estos sistemas multimodales, a fin, de ampliar el panorama de utilidad de los mismos.

## 1.6 Ventajas y desventajas de un sistema multimodal

La dinámica de los sistemas multimodales, se encuentra actualmente de una evolución y crecimiento constante. En este sentido, cabe señalar que existen una serie de ventajas y desventajas que se presentan para los distintos contextos en donde estos sistemas son empleados. Por ello, se hace necesario hacer una descripción de las fortalezas y debilidades de estos sistemas para evaluar la factibilidad de aplicación de éstos sistemas.

### 1.6.1 Ventajas

Las características que han hecho especialmente atractivos a los sistemas multimodales son básicamente las siguientes:

#### 1.6.1.1 Libertad de tareas simultáneas

La utilización de sistemas multimodales, permite la utilización de entrada de datos por medio de diferentes medios, de los cuales, se puede aprovechar la independencia de tareas. Esto implica, que el sistema puede ser utilizado mientras se puede llevar a cabo alguna otra tarea de forma simultánea.

Si bien es cierto, que bajo el concepto de multimodalidad, se ha definido la combinación de varias modalidades para entrada de datos en un sistema, estas modalidades se enfocan en medios naturales de comunicación. Estos medios de comunicación naturales, permiten realizar tareas simultáneas de igual forma que en una interacción entre seres humanos.

Por ejemplo, un sistema multimodal con una interfaz telefónica por medio de la voz, puede consistir en mantener las manos libres y en algunas ocasiones hasta la vista. En un sistema multimodal de autenticación biométrica, se puede seguir conversando por medio de la voz, mientras se realiza la verificación de una huella digital y la retina del ojo. Es importante la libertad de tareas simultáneas de un sistema multimodal, debido a que es una característica que se puede explotar a favor de la interacción de usuarios, permitiendo flexibilidad y aumentando la productividad y optimización de tareas.

#### 1.6.1.2 Utilización de medio naturales de comunicación

La interacción entre seres humanos es muy compleja. La conversación entre 2 o más personas puede conllevar la utilización de muchos medios de comunicación, por ejemplo, la utilización de voz, de señas, de gestos, etc. De igual forma, se puede reconocer personas conocidas por medio de la voz, si se le escucha por teléfono o de la vista, si se le ve por la calle.

Estas circunstancias y escenarios son bastante comunes para el ser humano. La utilización de medios naturales de comunicación en los sistemas multimodales, como reconocimiento de voz, de gestos, de movimientos por medio de un guante virtual, de la retina de ojo, etc., permiten que un intercambio de ideas o acciones entre seres humanos, se pueda trasladar a un sistema multimodal en una interfaz humano-computadora.

La introducción de estas acciones entre seres humanos a una interacción de una interfaz humano-computadora, es de gran utilidad para los usuarios, ya que de alguna forma se puede “humanizar” ésta relación, con lo cual, se brinda un entorno familiar para cualquier usuario que utilice el sistema.

#### 1.6.1.3 Unificación de tareas

El concepto de multimodalidad, que describe la utilización de más de una tarea en conjunto, es importante y valioso para potenciar la entrada de datos para un sistema. Esto se ve reflejado en la utilización de todas las ventajas de las tareas individuales, que al unificarse y combinarse, producen una tarea que es más fuerte que cualquiera de las individuales.

Sin embargo, se debe analizar el tipo de unificación que se pueda realizar entre las modalidades. Como se ha expuesto anteriormente, la utilización de características humanas, permite el análisis de la interacción entre personas y derivado de esto, se realiza el estudio de la combinación de modalidades que ya se utilizan en la interacción de seres humanos. Este estudio, se realiza en función de la facilidad que se le brinde al usuario, así como también, que las entradas del sistema resulten fortalecidas por la combinación de ventajas de modalidades individuales. Es mas fácil, para una persona que quiere transmitir un mensaje, el realizar esta acción hablando y apuntando, que sólo hablando o sólo apuntando.

#### 1.6.1.4 Interacción con el usuario

Los sistemas multimodales, permiten una interacción cómoda con el usuario, debido a que la interfaz se realiza con muchas de las características propias de la comunicación humana. El objetivo de la interfaz de usuario de cualquier sistema, es ofrecer al usuario, un entorno manejable y amigable. Bajo ésta premisa, el entorno que ofrecen los sistemas multimodales, al basarse en particularidades propias de la comunicación humana, proponen una opción importante en la medida de que el usuario se pueda adaptar a una situación propia de humanos, trasladada a una situación de humanos con un computador.

Esta interacción con el usuario, se ve beneficiada de igual manera, en la entrada de datos al sistema, ya que una persona puede ingresar u obtener información de forma natural. Al utilizar medios como la voz, los gestos, una huella digital, la retina del ojo, el tacto, etc., se puede ayudar a un usuario a interactuar con el sistema con sus características propias.

La interacción se puede realizar, incluso, sin necesidad de utilización de periféricos como un teclado o un ratón. Un ejemplo de esto, es el reconocimiento de voz, al unificar el reconocimiento de labios y la voz en un ambiente ruidoso. La facilidad para el usuario es que solamente necesitará hablar y el sistema, debido al entorno ruidoso, utilizara el reconocimiento de voz y de labios para entender el mensaje.

#### 1.6.1.5 Accesibilidad en diferentes contextos

La accesibilidad a los sistemas de información, debe extenderse a todos los niveles. Algunas veces, los sistemas se orientan hacia un grupo determinado. Sin embargo, existen grupos de usuarios que no tienen acceso a estos sistemas.

En la actualidad, el poseer conocimiento del manejo de una computadora, es básico para el desempeño de las labores diarias. No obstante, existen grupos de usuarios que no involucran el uso de una computadora en su trabajo u hogar. Este grupo de usuarios neófitos en este tema, pueden acceder a estos sistemas multimodales, ya que al utilizar las características humanas de interacción no se necesita que posean conocimientos del uso de un computador.

Así mismo, existen grupos de usuarios que debido a algún impedimento físico no se les permite interactuar con un computador. Al utilizar sistemas multimodales, se puede potenciar otras modalidades para que personas con impedimentos físicos puedan acceder a información y puedan utilizar la interfaz. Por ejemplo, personas no videntes pueden utilizar una interfaz que pueda reconocer la voz y reconocer el tacto, como se presentó anteriormente con el “WebTouch”.

Con los sistemas multimodales, se puede integrar más grupos de usuarios, en diversos entornos y que estos sean beneficiados por la interacción de varias modalidades.

### 1.6.2 Desventajas

Al igual que existen ventajas con los sistemas multimodales, se tienen ciertas desventajas que se presentan a continuación:

#### 1.6.2.1 Complejidad de implementación

Los sistemas multimodales, proveen una ventaja en la unificación de modalidades, pero de igual forma, esto implica una complejidad en la implementación. Al utilizar características de la comunicación entre seres humanos en sus interfaces, los algoritmos para desarrollar estas técnicas tienden a ser complicados. Si a esto se agrega, que se le unificará otra modalidad, con algoritmos igualmente dificultosos, el resultado es una aplicación con un nivel de complejidad alto.

Este nivel de complejidad depende de la aplicación que se desarrolle y de las modalidades que se unifiquen. El fortalecimiento de la unificación de modalidades tiene un costo en la implementación.

#### 1.6.2.2 Manejo de errores dificultoso

La unificación de tareas en un sistema multimodal, hace que la implementación sea compleja y por ello, el manejo de errores sea dificultoso. La complejidad de los algoritmos y la combinación de los mismos, trae como consecuencia, la probable dificultad en el control, detección y eliminación de

errores. En algunas modalidades de la comunicación entre seres humanos, aun se tienen varios aspectos por corregir, aun no se cuenta con algoritmos que logren el 100% de confiabilidad respecto a reconocimiento de patrones humanos.

Esto puede estar presente en cada una de las modalidades combinadas, con lo cual, se proporcionaría un margen de error que no permite una pronta y correcta detección de errores, ya que al no tener el 100% de confiabilidad, el error puede ocurrir en ese margen creando ambigüedad en su procedencia.

#### 1.6.2.3 Aprendizaje del comportamiento del usuario

Los algoritmos para reconocimiento de patrones humanos, utilizan técnicas de enseñanza conforme se adaptan al medio. Por lo tanto, para que el sistema funcione de forma adecuada, debe pasar inicialmente por una etapa de aprendizaje, tanto de los usuarios que lo utilizarán, como del medio ambiente que lo rodea.

Esta etapa de aprendizaje debe realizarse sobre un conjunto de escenarios y usuarios, a fin de que la aplicación pueda adaptarse y aprender las diversas situaciones que puede encontrarse. Esto implica tiempo en el desarrollo del aprendizaje, así como la capacidad de la conjunción de modalidades de asimilar el entorno y potenciales usuarios del sistema.

#### 1.6.2.4 Ambientación a la interfaz por parte del usuario

Del lado del usuario, el utilizar una interfaz humano-computadora con una interacción “humanizada” (el utilizar la voz, gestos, etc) no es del todo familiar. La ambientación del usuario a la interfaz, es importante para el buen manejo del

sistema. Aquí se debe romper con la idea que, una computadora no puede realizar una interacción con un ser humano como si fuera otro ser humano.

De hecho, no es que una computadora se transforme en un ser humano, sino que las entradas de datos sean mas cómodas para el usuario proporcionando características propias del humano para esa entrada de datos. Esta ambientación lleva tiempo y entrenamiento para el uso adecuado de la interfaz, lo cual tiene un costo en la etapa de aprendizaje y pruebas del sistema.

Los sistemas multimodales, son aplicables en muchos ámbitos de la vida real, tanto para interfaces humano-computadora, como para sistemas donde no se involucre una máquina, como se ha expuesto en este capitulo. De igual forma, se han definido los conceptos relacionados a esta metodología, los componentes que se utilizan en las interfaces hombre-maquina, los puntos positivos y negativos de estos sistemas. Todo esto se presenta, con el fin de especificar el área de trabajo de la multimodalidad, independientemente de la interfaz que se utilice. Con esto, se puede observar el campo de acción de estos sistemas y la aplicación de los mismos en un contexto general.

# 2. MODALIDAD DE VOZ

Entre las formas de comunicación más comunes que existen, se encuentra la voz. La voz, es una forma natural de comunicación que es persuasiva, eficiente y puede ser usada a distancia. Este tipo de comunicación, es una manera rápida de expresar ideas para cualquier hablante, debido a que se aprende desde el mismo entorno. Hay muchos conceptos asociados a esta modalidad de voz y al proceso que la involucra, es decir, el proceso de comunicación. Estos conceptos se detallan a continuación para entender en todo sentido la importancia de la voz en el ser humano. Un sistema multimodal permite la interacción con el usuario a través de diferentes canales de interacción como la voz, gestos, escritura en un teclado o uso de un ratón en una computadora, con el fin de facilitar la obtención del significado de la interacción del usuario.

## 2.1 Sonidos, palabras, el habla y la lengua

El elemento básico, para que la comunicación por medio de voz se produzca, son los sonidos. El sonido, es un elemento indispensable en este proceso de comunicación. Es una onda mecánica longitudinal que se propaga a través de un medio material (sólido, líquido o gaseoso). Este fenómeno sonoro, esta relacionado con las vibraciones de los cuerpos materiales. Siempre que se escucha un sonido, hay un cuerpo material que vibra y produce ese fenómeno. Para producir los sonidos del habla se utilizan los labios, dientes, lengua, cavidades nasales, cuerdas bucales y sistema respiratorio. Cuando una persona habla, el sonido que emite es producido por la vibración de sus cuerdas vocales, las cuales producen ondas que se

propagan en el medio material situado entre ellas y el oído. Al penetrar en el órgano auditivo, estas ondas producen vibraciones que causan las sensaciones sonoras. El oído humano, en situaciones normales, puede captar sonidos de una frecuencia entre 16 y 20,000 ciclos por segundo (vibraciones dobles por segundo o hertz), aunque por lo general, es más sensible a las diferencias entre un tono y otro cuando se hallan a 50 dB (decibeles) por encima del umbral de audición y en la gama de los 500 a 4,000 ciclos por segundo (zona de la discriminación auditiva del sonido). Una persona puede percibir sonidos entre 0 y 120 dB (decibeles), es decir, entre el mínimo nivel posible de detectarlos y el umbral de la molestia.

Luego de analizar la unidad básica de la voz, los sonidos, se da paso a la combinación de éstos, dando lugar a las palabras. Las palabras, son un conjunto de sonidos articulados que expresan una idea. Las palabras a su vez, en conjunto, dada la unión de éstas, originan las frases. Pero este proceso de combinar sonidos para obtener palabras y en conjunto, las palabras generar frases, se rige por una serie de reglas lingüísticas muy complejas y variadas (fonéticas, sintácticas o semánticas) que constituye cada uno de los sistemas de cada lengua, el cual, se materializa en el habla concreta de cada individuo de una comunidad lingüística.

El habla, es la producción de sonidos con voz y articulación para comunicar los pensamientos de nuestro cerebro y liberar emociones. En principio, se puede adquirir los específicos sonidos del habla a los que se está expuesto, los que se escuchan en el ambiente. Los sonidos que no se escuchan, difícilmente se aprenden, de hecho, las personas aprenden a escuchar los sonidos. El habla y el oído están íntimamente relacionados. Este vínculo es inconsciente, no se tiene que pensar concientemente en dónde colocar la lengua para poder hablar. Pocas personas pueden decir donde

colocar la lengua para producir el sonido "R", aunque toda la gente lo hace cientos de veces al día. Ni siquiera se tiene que pensar para nada a cerca de la respiración, cuando se pasa la comida, cuando alguien habla y cuando alguien se ríe, es totalmente automático. El mecanismo del habla materializa la lengua.

La lengua, es el código de la comunicación lingüística. Se aprende sin que las personas se den cuenta. Conceptualmente, la lengua se define como un conjunto de imposiciones, pero también, y quizá mejor, como conjunto de libertades, puesto que admite infinitas realizaciones y sólo exige que no se afecten las condiciones funcionales del instrumento lingüístico[16].

Cuando las personas hablan, comunicándose sus pensamientos, sus ideas, comprendiéndose entre si, es porque existe algo común y que está en un plano superior a ellos mismos, es decir, se entienden porque existe la lengua, el modelo lingüístico común a los dos, el sistema que establece ciertas reglas a las que se someten cuando hablan; y en el momento en que expresan sus ideas oralmente, están, materializando la lengua en cada uno de ellos, están practicando el acto de habla. El plano de la lengua y el plano del habla están ligados estrechamente, unidos inseparablemente y constituyen los dos aspectos del fenómeno conocido con el nombre del lenguaje, es decir, la capacidad que tiene el hombre para comunicarse con los demás por medio de signos orales o escritos.

[16] Referencia bibliográfica [5] Sistemas Alternativos de Comunicación.

## 2.2 Lenguaje oral y hablado

El lenguaje oral, es la forma plenaria de la comunicación a la que, en mayor o menor medida, se incorporan todas las demás. El lenguaje, desde la perspectiva de la psicología popular, y al igual que ocurre con el término de la comunicación, es un concepto de límites difusos, con muy diferentes significados y de muy diferentes niveles de complejidad en cada una de ellas. Tomando una de las múltiples definiciones que se han dado sobre el lenguaje, según Baker y Cokely lo definen así:

"Un lenguaje, es un sistema de símbolos relativamente arbitrarios y de reglas gramaticales que se transforman en el tiempo y que los miembros de una comunidad convienen y usan para interactuar unos con otros, comunicar sus ideas, emociones e intenciones, para pensar y para transmitir su cultura de generación en generación"[17].

El lenguaje, también se puede definir como la facultad de la mente humana con base en la cual se codifica o descodifica un mensaje, operacionalizada mediante una estructura neuropsicológica que está conformada por una red de alta complejidad de mecanismos y centros nerviosos especializados genéticamente en la organización de la producción y el reconocimiento de los sonidos, las reglas que gobiernan el ordenamiento secuencial de las palabras en frases y oraciones; y el sistema de significado que se adhiere a éstas como consecuencia de las experiencias cotidianas y la interacción social del individuo en una variedad de situaciones comunicativas, desarrollados durante la evolución del hombre.

[17] Referencia bibliográfica [5] Sistemas Alternativos de Comunicación.

En un sentido más restringido, más de estudio psicológico del lenguaje, éste se caracteriza por un sistema estructurado, complejo, flexible y convencionalizado de elementos, que sirven para representar aspectos de la realidad distintos de los elementos de los mismos del sistema, y para llevar a cabo actos de comunicación.

Pero el lenguaje oral o hablado, es algo más que producir una serie de sonidos o gestos para crear palabras o signos; es también un sistema de reglas a otros niveles (reglas morfológicas y sintácticas), que deben ser conocidas y compartidas por todos lo que comunican esa lengua. Todas las lenguas tienen sus reglas morfosintácticas. Se dice por tanto, que es un sistema convencional y arbitrario con las funciones de comunicar y representar. El lenguaje, antes que nada, es combinación y construcción de la identidad social y personal. Es la forma de comunicación verbal principal, que pueden usar todos los hablantes, cualquiera que sea su nivel sociocultural. Es también la forma viva y espontánea de la lengua.

Los rasgos que caracterizan genéricamente al lenguaje hablado son su uso utilitario y su propósito de comunicación. Nunca se escribe, exactamente igual que como se habla, ya que la lengua hablada permite palabras, construcciones, interrupciones, incorrecciones y desórdenes que no son posibles ni permisibles en la lengua escrita. Mejor o peor, con mayor o menor propiedad, todo el mundo sabe hablar para entenderse con los demás, esto es, para comunicar algo, expresar lo que le sucede o siente o actuar sobre el interlocutor. Estas son en síntesis, las funciones del lenguaje.

Al hablar, se cometen abundantes incorrecciones, se utilizan escasas palabras, muchas veces ni siquiera se rematan las frases, y es porque los

gestos, la expresión del rostro, el tono de voz, la situación en que se habla, contribuyen a entender y ser entendido.

El lenguaje oral, es parte de la comunicación oral o hablada, en el cual se involucran varias fases y varios elementos. Es un proceso fundamental para el ser humano pueda satisfacer su necesidad de comunicar sus carencias e ideas.

## 2.3 Proceso de comunicación oral y hablada

En un mundo en que la necesidad de la relación con los demás se manifiesta a través de todos los niveles y en las actividades más diversas, el tema de la comunicación ha adquirido un extraordinario relieve. Sin las palabras y la capacidad de expresarlas por medio de la voz, resultaría en extremo difícil y casi imposible, coordinar las actividades más elementales de la vida en relación con los demás. Cuando un ser humano comunica sus necesidades e ideas a otros para lograr su comprensión o conseguir su cooperación, debe hablar bien, de modo coherente, convincente y preciso.

La comunicación es dinámica, un proceso en movimiento y sirve para realizar distintas funciones. Desde muy temprano un niño empieza a descubrir conexiones. Los gritos sirven para atraer comida y atenciones. Dentro de la familia, el niño empieza a conocer los sistemas de cooperación y competencia; también aprende las formas aceptables y no aceptables de comunicación.

Puede examinarse a la comunicación desde el punto de vista de la construcción del mensaje. Esto significa que puede tomarse toda la información necesaria para presentarse en una situación dada, respecto a todas las funciones o variables de la comunicación, y estructurar un mensaje. Se debe considerar al mensaje como un código, o como un sistema de comunicación

que pasa a través de muchos niveles de extracción, se adapta a muchas variables y por lo tanto se obtienen respuestas según las unidades que participan. Los componentes básicos de un sistema de comunicación son: la unidad receptora, la unidad procesadora y la unidad transmisora.

En un sentido amplio, se puede definir la comunicación como el acto mediante el cual una persona o personas transmiten a otra u otras, y por cualquier procedimiento, mensajes de contenido diverso, utilizando intencionadamente signos dotados de sentido para ambas partes y que establece una relación que produce unos efectos. Una definición más esquemática seria la de transmisión de un mensaje entre un emisor y un receptor. En cualquiera de las dos están implícitos los componentes básicos anteriormente mencionados.

### 2.3.1 El emisor y receptor

El emisor (individual o colectivo) es el que emite el mensaje. Elige y selecciona las señales que le convienen, es decir, realiza un proceso de codificación (representa en un código lo que va a transmitir). A esta operación de codificación le acompaña una intención (el emisor pretende comunicar algo concreto con un fin determinado). Los datos que transmite el emisor proceden de una fuente de información. En el caso de una conversación interpersonal, de los conocimientos o sentimientos del emisor; en el caso de una comunicación social, la procedencia o el hecho que se desea transmitir.

El receptor (individual o colectivo) es el destinatario del mensaje. Interpreta dicho mensaje, esto es, lo descodifica (descifra el código empleado por el receptor). El proceso de descodificación se efectúa en el nivel receptor-destinatario por medio de la búsqueda en la memoria de los elementos

pertenecientes al código que se ha seleccionado para la trascripción del mensaje. La actividad del receptor dentro del acto comunicativo depende de tres factores:

- De la recepción misma, condicionada por la atención que el destinatario presta al mensaje.
- De su participación al atribuir significados a los signos recibidos y sentido a los mensajes emitidos. En eso consiste la operación de descodificar e interpretar los mensajes que se reciben, que sitúan al destinatario en una posición activa y colaboradora. Si el mensaje no llega a descodificarse e interpretarse no hay comunicación aunque haya percepción de señal (imagen, sonido, escrito…). Por eso los idiomas suelen distinguir entre mirar y ver, y entre oír y escuchar.
- Del consecuente efecto interno (sea en el conocimiento, la experiencia, la sensibilidad, la actitud) o externo (opinión y conducta).

### 2.3.2 El mensaje

El mensaje es la secuencia (oral o escrita, verbal o no verbal) de elementos tomados de un repertorio de signos por el emisor para transmitirlos al receptor. Para los teóricos de la comunicación este término no significa más que una secuencia de señales transmitidas entre un emisor y un receptor a través de un canal que sirve de soporte físico a la transmisión.

El mensaje es el algo que comunicar, el contenido, compuesto o cifrado por el emisor ajustándose al código. Para transmitir un mensaje es necesario codificar la información. Esa codificación puede realizarse de forma gestual, verbal, visual, audiovisual, etc.

### 2.3.3 El código

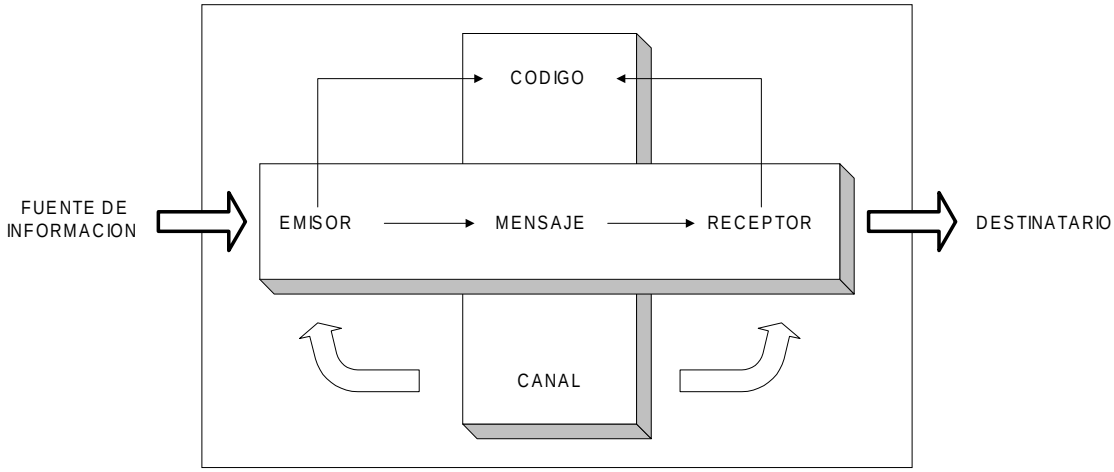
El código es el conjunto de signos y reglas que se emplean y combinan. Un código puede tener un número muy reducido de señales y reglas de combinación, o por el contrario, un número muy elevado; puede ser común a un gran número de emisores y receptores o estar restringido a un pequeño número (dos como mínimo). Se trata siempre de una potencialidad y su única manifestación posible es en forma de mensaje.

A veces no basta con que el código lingüístico, sea el mismo para el emisor y el receptor; es necesario también que coincidan los sub-códigos. El emisor y el receptor deben tener un conocimiento amplio y rico del código general o de los sub-códigos correspondientes.

### 2.3.4 El proceso

Analizados los elementos que intervienen en el acto de la comunicación, veamos como se produce el proceso. En el esquema mas simplificado de la comunicación, intervienen, como mínimo, un emisor, un mensaje y un receptor. El objetivo de la relación determina los papeles que los participantes juegan en este proceso. El proceso, con sus protagonistas, se ejemplifica con la siguiente figura:

**Figura 1. Proceso de comunicación**




El emisor envía un mensaje al receptor. Previamente, este emisor, utilizando sus necesidades y recursos cognitivos, así como sus habilidades para comunicarse, codifica ciertos signos de información y los ofrece al receptor, que los interpreta según su propio mapa cognitivo, añadiéndole, además, algunos rasgos informales e involuntarios del emisor como su expresión facial, su interés, su desinterés, etc. Para que sea operativo, el mensaje requiere previamente un contexto al que remite (referente), contexto comprensible para el emisor. Luego, el mensaje necesita un contacto con el canal físico, que le permita establecer y mantener la comunicación.

Por otra parte, los papeles del emisor y del receptor son reversibles: las mismas personas los interpretan alternativamente, por ejemplo, en la conversación, que es el prototipo de la comunicación lingüística. Emisor y receptor funcionan de forma interactiva, es decir, el emisor pasa a ser receptor de un nuevo mensaje. Este proceso dinámico se da mientras dure la conversación. La respuesta al mensaje se llama retroalimentación de la comunicación.

La comprensión de un mensaje, informa hasta que punto ha sido codificado correcta o incorrectamente, y puede ofrecer señales que indiquen que se debe emitir desde otra perspectiva si queremos que sea recibida. La retroalimentación puede adoptar la forma de preguntas, gestos, respuestas, etc. Permite corregir las posibles omisiones y errores en la transmisión de un mensaje, o mejorar la codificación y el proceso de transmisión.

### 2.3.5 Objetivos de la comunicación oral

Cuando una persona habla, es decir, cuando se comunica con alguien a través de la palabra, intenta lograr los siguientes propósitos:

#### 2.3.5.1 Concreción de la idea

Es de vital importancia saber lo que se que quiere decir exactamente, para que se logre una comunicación adecuada. Algunas veces, los seres humanos divagan en sus ideas expresándolas, de tal manera que, a medida que hablan se forma la idea que se quiere expresar y si la ocasión se presenta, se amplia, se rectifica, se explica la idea de otra forma. En este rodeo de palabras, se pierde tiempo y atención por parte del receptor. El tener claro la idea que se expresará, de manera directa, provocará que el receptor tenga menos problemas en la comprensión del mensaje. Se trata, en otras palabras, de decir algo que ya haya tomado forma previamente en nosotros, de expresar una idea concreta.

#### 2.3.5.2 Adecuación del tono

Se debe encontrar el tono adecuado para lograr que el receptor o destinatario acepte el mensaje y consecuentemente, la traduzca en acción.

Aun cuando se tenga la idea concretada, perfectamente elaborada, se debe considerar ese elemento sonoro no verbal, que es el tono. El tono, no es más que un regulador entre el sentimiento y la expresión, entre lo que se siente y lo que se dice. Por consiguiente, es fundamental para lograr una buena comunicación, encontrar el tono adecuado, es decir, no actuar solo sobre las palabras midiéndolas y pesándolas, sino ir al fondo del problema.

#### 2.3.5.3 Coordinación del mensaje

El emisor del mensaje no puede perder nunca el contacto con el oyente, porque se corre el riesgo de que cese la atención y de que el mensaje no llegue, o llegue de manera imperfecta al receptor. El receptor debe estar en condiciones de seguir al emisor en su mensaje, de manera que pueda ir asimilando y entendiendo cada palabra. Para lograr esto, el emisor debe colocar las ideas una después de las otras, coordinándolas de manera tal, que no se pierda el mensaje. La sencillez y la coordinación de las ideas que forman el mensaje es esencial para una buena comunicación.

#### 2.3.5.4 Utilización de términos exactos

Es posible, que se esté en posesión de la idea concreta, que se esté usando el tono más adecuado para llegar al receptor, y que esté siguiendo paso a paso, entendiendo y asimilando, cuanto se le transmite. Aun así, puede ocurrir que en un momento determinado no se encuentre la palabra exacta para la expresión de la idea. Hay una clave para encontrar la palabra exacta, que no ha de ser la palabra precisa, sentir, vivir, ver y comprender aquello que se está hablando. Lo importante en este sentido, es no perder de vista la idea o la realidad de que se esté hablando.

### 2.3.6 Cualidades del estilo oral

El estilo, es la manera propia que cada persona tiene para expresar su pensamiento por medio de la escritura o de la palabra. Es la manera que cada individuo tiene de crear expresiones para comunicar su pensamiento. Las cualidades primordiales del estilo oral son las siguientes:

#### 2.3.6.1 Claridad

En términos generales, claridad significa expresión al alcance de un hombre de cultura media, pero quiere decir además, pensamiento transparente, conceptos bien digeridos, exposición limpia, es decir, con sintaxis correcta y vocabulario o léxico al alcance de la mayoría. Dicho de otro modo, un estilo es claro cuando el pensamiento del que emite el mensaje penetra sin esfuerzo en la mente del receptor.

#### 2.3.6.2 Concisión

La concisión resulta de utilizar sólo palabras indispensables, justas y significativas para expresar lo que se quiere decir. No debe entenderse, sin embargo, que estilo conciso sea sinónimo de estilo breve y ultracondensado, sino que la concisión es enemiga del palabreo, de la redundancia, del titubeo expresivo, porque todo esto obstruye los canales de la comunicación y el mensaje no llega adecuadamente al receptor.

#### 2.3.6.3 Coherencia

Cuando se habla, cuando se comunica oralmente, el orden en el correr de las ideas ha de ser tal, que el oyente no se vea precisado a coordinarlas en su

cerebro. Las relaciones entre las ideas expuestas deben ser lógicas, y las contradicciones evitadas. La coherencia es la conexión, relación o unión de ideas con otras.

#### 2.3.6.4 Sencillez

La sencillez, es otra condición o cualidad necesaria del buen estilo en la comunicación oral, que se refiere tanto a la composición de lo que hablamos, como a las palabras que empleamos. Ser sencillo no es, sin embargo, tan fácil como pudiera creerse. Se debe buscar las palabras adecuadas y simples para expresarse de la mejor manera, sin llegar a complicaciones.

## 2.4 Ventajas y desventajas de la comunicación oral

La importancia de la comunicación oral en la contexto humano es fundamental, ya que es la forma más común y natural de expresar ideas. Debido a esto, es importante conocer los factores positivos y negativos que se puedan presentar. A continuación se realiza una descripción de las fortalezas y debilidades de esta forma de comunicación.

### 2.4.1 Ventajas

La comunicación hablada presenta beneficios en su forma de transmitir el mensaje a receptores, beneficios que se presentan a continuación.

#### 2.4.1.1 Utilización generalizada por las personas

La comunicación hablada, es la forma más utilizada por las personas para expresar sus ideas. La mayoría de la gente, aprende a hablar un lenguaje

específico dependiendo del ambiente en que se desenvuelva, independientemente del nivel sociocultural que se maneje. Por ejemplo, en América Latina, el lenguaje que comúnmente se adquiere es el español, debido que es el que predomina en el ambiente.

El hecho de que la mayoría de las personas utilizan la comunicación oral, es un factor determinante, ya que es innato en cualquier ser humano, la propiedad de hablar y comunicarse de esta manera, no hay que tener ningún tipo de adiestramiento en este aspecto.

#### 2.4.1.2 Continua evolución

La comunicación oral, es un elemento que constantemente se desarrolla en tal forma que el hablante puede tener mayor libertar para expresarse. Este aspecto concede al emisor un amplio margen de libertad para hablar, ya que cada persona tiene su propia manera de expresarse y de hablar, con las palabras que crea conveniente.

#### 2.4.1.3 Adquisición de forma natural

El hablar es algo natural, desde que un niño empieza a balbucear, es señal inequívoca que esta aprendiendo del ambiente en que se encuentra, concretamente el lenguaje y los sonidos que percibe. Regularmente las personas aprenden el lenguaje como algo natural y espontáneo, aprendiendo del ambiente. Nadie le dice a un niño que empieza a decir sus primeras palabras como tiene que colocar la lengua para expresar un sonido que lleve luego a una palabra, esto es algo natural.

#### 2.4.1.4 Manifestación en una situación concreta

Cuando se entabla una comunicación hablada, normalmente el hablante y el oyente se encuentran en ocasiones en el mismo lugar, tiempo o circunstancias. Esto hace que la comunicación sea mejor, ya que, al estar en el mismo lugar, tiempo o circunstancia el mensaje se puede captar de una manera adecuada.

### 2.4.2 Desventajas

A fin de tener el contexto general de la comunicación hablada, se presentan los puntos que pueden considerarse como desventajas, lo cuales básicamente son los siguientes:

#### 2.4.2.1 Ruidos

El proceso de comunicación hablada puede sufrir diversas distorsiones a causa de defectos de los aparatos receptores o emisores, de perturbaciones en el canal, etc. Estas distorsiones influyen en la calidad del mensaje transmitido, una parte del cual puede perderse durante dicho proceso. Todas estas perturbaciones se denominan ruidos. El ruido puede llegar a: formar parte de la información, deteriorarla o destruirla. Es un aspecto el cual, se encuentra en el ambiente en cualquier momento.

#### 2.4.2.2 Sustitución de sonidos por gestos

Cuando surge una comunicación hablada, los que intervienen en el proceso pueden utilizar gestos para sustituir palabras. Pueden dejar frases inconclusas y terminarlas con gestos. Esto es parte del proceso de

comunicación, pero afecta directamente a la comunicación hablada. El factor que entra a jugar en contra de la comunicación oral es el contacto visual, el percibir comunicación por medio de la vista.

#### 2.4.2.3 Códigos Lingüísticos

Existe un problema en los códigos lingüísticos que se utilicen en una conversación hablada. Existen muchos códigos lingüísticos por cada región del planeta y no hay una forma de unificarlos por medio de la voz. Si los códigos lingüísticos son distintos en una comunicación oral, será muy difícil que este proceso se lleve a cabo con éxito.

#### 2.4.2.4 Modismos

Al igual que la comunicación oral evoluciona continuamente, también existen modismos o sub-lenguajes los cuales no entran en las reglas lingüísticas establecidas por ninguno de los lenguajes. A estos sub-lenguajes se les llama caló. Estos son una sub-utilización de los lenguajes existentes, haciendo modificaciones a palabras y significados de las mismas.

En síntesis, la voz, es una forma natural de comunicación, por medio de la cual, los seres humanos suplen la necesidad de relacionarse con los demás. Es el medio más importante de comunicación y mas utilizado por la gente para comunicarse. La importancia de la voz en el quehacer humano es trascendental en el diario vivir de éste mundo.

# 3. TECNOLOGÍA DE RECONOCIMENTO DE VOZ

El objetivo de la tecnología de reconocimiento de voz, en un sentido amplio, es crear máquinas que puedan recibir información hablada y actuar apropiadamente con esa información; además del intercambio de información de la máquina al humano usando voz sintética. En estos términos, el estudio de reconocimiento de voz es parte de la búsqueda de máquinas "artificialmente inteligentes" que puedan "oír", "entender", y "actuar con" información hablada, así como "hablar" para completar el intercambio de información.

La investigación de tecnologías en reconocimiento de voz empezó a finales de los 50´s con la llegada del computador digital. Esto combinado con herramientas para capturar y analizar la voz permitió a los investigadores encontrar otras maneras de representar características acústicas que mostraran las diferentes propiedades de las palabras. Uno de los pioneros en este campo fue AT&T®. El sistema desarrollado por esta compañía se entrenó para reconocer el discurso de manera dependiente del locutor.

En la época de los 60´s la segmentación automática de voz logró avanzar en unidades lingüísticas relevantes (fonemas, palabras y sílabas), clasificación y reconocimiento de patrones. Inicialmente los investigadores habían subestimado la dificultad de la tarea, sin embargo pronto para simplificar sus planes y las aplicaciones que se desarrollaron, se fueron haciendo dependientes del locutor y con vocabularios pequeños.

En los 70´s surgieron un número de técnicas realizadas en su mayoría por la Agencia DARPA (Defense Advanced Research Projects Agency, por sus siglas en inglés). Se desarrollaron reconocedores que manejaban un dominio

de reconocimiento mayor basados en el reconocimiento de patrones. Los reconocedores eran capaces de aceptar un vocabulario más extenso. Durante esta época se logró una mejora con respecto al reconocimiento para palabras aisladas y continuas.

Los 80´s se caracterizaron por el fuerte avance que se obtuvo en el reconocimiento de voz. Se empezaron a desarrollar aplicaciones con vocabularios grandes y se impulsaron el uso de modelos probabilísticos y redes neuronales, los cuales poco a poco mejoraron su desempeño.

Para los 90´s el progreso de los sistemas de reconocimiento de voz es notable gracias a la innovación de la tecnología (computadoras y algoritmos). Los investigadores realizaron vocabularios grandes para usarse en el entrenamiento, desarrollo y pruebas de los sistemas. Además, las técnicas de hace algunos años han sido mejoradas para obtener mejores reconocedores.

En resumen, el reconocimiento de voz es el proceso de convertir una señal acústica, capturada por un micrófono o teléfono, a un conjunto de palabras. Las palabras reconocidas pueden ser el resultado final para aplicaciones de comandos y control, entrada de datos y preparación de documentos.

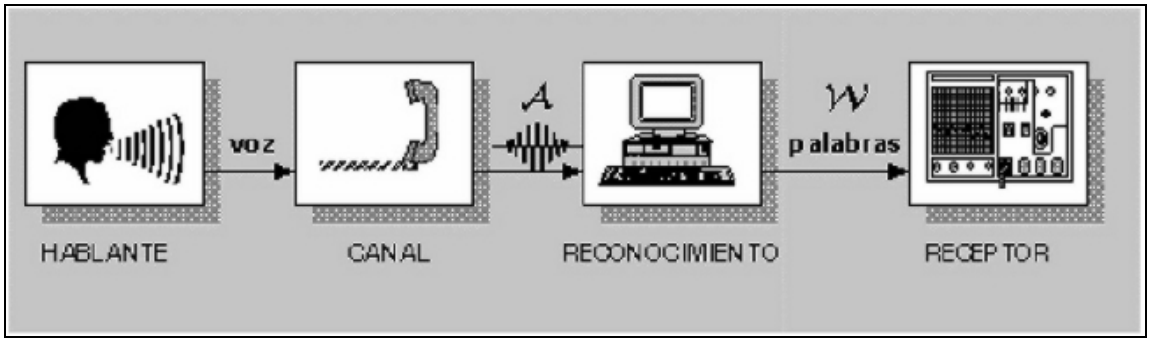
## 3.1 Componentes

Un Sistema de Reconocimiento de Voz consiste de los siguientes elementos:

- Un canal de entrada, que puede ser un micrófono o una línea telefónica, que recibe la voz correspondiente a la pronunciación de una secuencia de palabras W.
- Un módulo de reconocimiento que al procesar la señal acústica de voz A produce la secuencia de palabras reconocidas W.
- Un módulo receptor que interpreta a W como un comando de control, un dato de entrada a una aplicación, o simplemente como el texto correspondiente al conjunto de palabras reconocidas.

En la figura 2 se presenta la descripción gráfica de un Sistema de Reconocimiento Automático de Voz que acepta como entrada una señal acústica de voz vía telefónica y produce como salida una secuencia de palabras que sirve de instrucción a una aplicación.

**Figura 2. Sistema de reconocimiento de voz automático**




### 3.2 Dominio de aplicación

Un acertado reconocimiento de las expresiones en lenguaje natural requiere de conocimiento acerca del mundo y acerca del contexto en el que se quiera trabajar e interactuar, a este último se le conoce como dominio de aplicación.

El éxito de un sistema de lenguaje hablado depende de su conocimiento del dominio de aplicación, es decir, que el conocimiento sobre el lenguaje que debe utilizar el sistema sea relativamente completo en relación a este dominio, y además de que los algoritmos del sistema sean apropiados y eficientes. Además, el desarrollo de un modelo de lenguaje natural aplicable a cualquier dominio parece ser imposible para todo propósito práctico. Por esa razón, aún se trabaja con dominios específicos.

Actualmente, los sistemas y herramientas computacionales que son capaces de manejar lenguaje natural adecuadamente, deben su éxito al uso de subconjuntos bien definidos de lenguaje natural, también llamados sub-lenguajes, es decir, se enfocan a un dominio de aplicación. Solamente a través de este tipo de compromiso, se han encontrado soluciones para varios problemas relativos al lenguaje. Esto implica que estos sistemas resuelvan tareas específicas y se enfoquen sobre dominios restringidos. Asimismo, se debe tomar en cuenta que en los diálogos orientados a tareas, la complejidad del diálogo depende de la complejidad de la tarea.

El procesamiento de lenguaje hablado utiliza conocimiento acerca de las palabras y de su pronunciación (léxico), además de un conjunto de reglas que definan la estructura correcta de las frases (gramática) con el objetivo de asignarles una estructura sintáctica. Generalmente, el léxico y la gramática corresponden a un dominio específico acorde a la aplicación en la que serán usados.

No existe demasiada información, sobre las características que debería tener un dominio de aplicación adecuado para implementar un sistema de comprensión del lenguaje hablado, que permita desarrollar la tecnología

necesaria y sobre todo, aprender todo lo necesario en este nuevo campo, encontrando respuestas y soluciones a tantas cuestiones como están planteadas. Sin embargo, en función de las razones o justificaciones dadas por distintos grupos de investigación que desarrollan su actividad en este tema, principalmente los participantes en *ARPA Spoken Language Understanding Systems*[18], se pueden resaltar los siguientes puntos:

- La aplicación debe ser lo más real posible, es decir, si existe una aplicación que esté siendo utilizada por un grupo muy numeroso de usuarios (y tenga un grupo potencial de usuarios muy importante), sin utilizar interfaces hombre-máquina capaces de entender el habla del usuario, mejor. Así podremos utilizar los datos y al grupo de usuarios de la misma. Además, existirá un interés comercial a posteriori si se consiguen resultados suficientemente buenos que justifiquen el uso de estas nuevas tecnologías.
- Los usuarios conocerán el uso de la aplicación y además, valorarán todo aquello que redunde en una mejora de las prestaciones que obtienen usando estas tecnologías, permitiendo capturar datos inicialmente y evaluar el comportamiento de los primeros sistemas implementados.
- El dominio debe ser suficientemente complejo para incorporar los elementos que pudiesen encontrarse en otras aplicaciones o dominios de naturaleza similar, es decir, nos debe permitir estudiar aspectos del diálogo entre el usuario y la aplicación, necesitar de interacciones habladas naturales, ricas en vocabulario, expresiones, poco restrictivas lingüísticamente hablando, donde los usuarios puedan comportarse con gran naturalidad, como si

[18] Referencia bibliográfica [33] Spoken language understanding

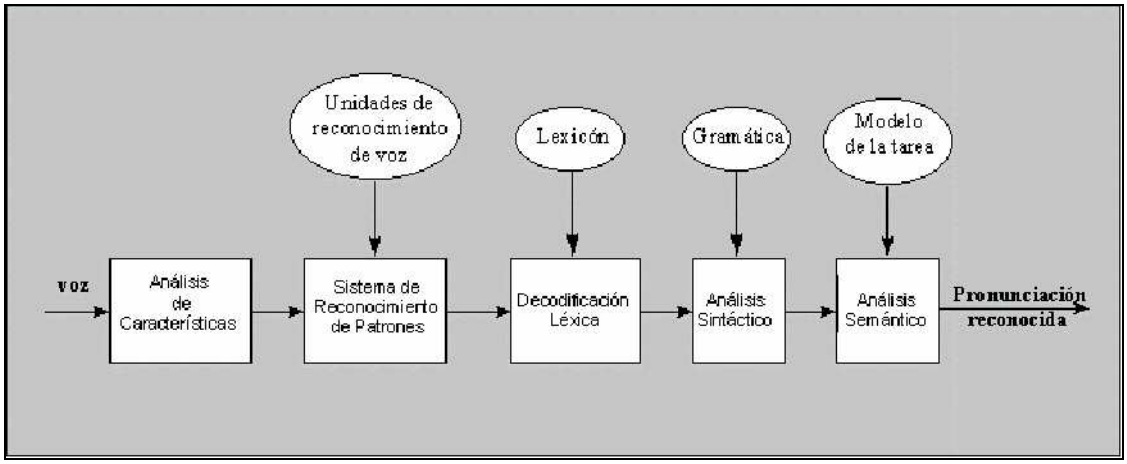
estuviesen hablando con un operador humano. Además, debe permitir reducir la complejidad de la aplicación inicialmente, manteniendo un alto grado de escalabilidad que permita ir aumentando las prestaciones del sistema de forma adecuada.

- El dominio debe ser suficientemente general, de modo que la tecnología desarrollada sirva para implementar aplicaciones en otros dominios, volviendo a utilizar la mayor parte del conocimiento obtenido.

## 3.3 Arquitectura

Un reconocedor de voz continua debe estar formado, por un módulo de análisis de características de la señal de voz, un sistema de reconocimiento de patrones, un módulo de decodificación léxica y un par de módulos más para análisis sintáctico y semántica. Se considera que los módulos de análisis sintáctico y semántica no pertenecen a la parte de reconocimiento de voz sino a la parte de procesamiento de lenguaje natural. De cualquier manera, en un sistema de lenguaje hablado deben estar incluidos este par de módulos para que el sistema actúe de acuerdo al dominio especificado, interprete correctamente las entradas y produzca los resultados apropiados. La figura 3 muestra la arquitectura de un sistema de reconocimiento de voz continua.

**Figura 3. Arquitectura de un sistema de reconocimiento de voz**




A continuación se describen los pasos clave involucrados en el reconocimiento de voz:

### 3.3.1 Análisis de características

En esta etapa se realiza un análisis espectral o temporal de la señal de voz para obtener los vectores de características que serán usados para entrenar a los modelos acústicos basados en redes neuronales o modelos ocultos de Markov que representan a los sonidos de voz.

### 3.3.2 Sistema de reconocimiento de patrones

En esta etapa se seleccionan las unidades de reconocimiento de voz que pueden ser unidades fonéticas, fonemas, bifonemas, sisabas o palabras. Cada unidad se caracteriza por algún tipo de modelo cuyos parámetros son estimados a partir de un conjunto de datos de voz de entrenamiento. El sistema de reconocimiento de patrones proporciona las probabilidades de la comparación de todas las secuencias de unidades de reconocimiento de voz

con la entrada de voz. Finalmente se determina la mejor aproximación sujeta a las restricciones léxicas y sintácticas del sistema.

### 3.3.3 Decodificación léxica

Este paso coloca las restricciones léxicas en el sistema de reconocimiento de patrones tal que las secuencias de palabras investigadas sean aquellas relacionadas a las secuencias de unidades de voz que están en el diccionario de pronunciaciones de palabras (léxico). Este procedimiento implica que el reconocimiento de voz del vocabulario debe estar especificado en términos de las unidades básicas seleccionadas para el reconocimiento. Cuando las unidades son palabras, el paso de decodificación léxica es prácticamente eliminado.

### 3.3.4 Análisis sintáctico

Esta fase coloca las restricciones en el sistema de reconocimiento de patrones tal que las secuencias de palabras investigadas son aquellas que están en una secuencia especificada por la gramática.

### 3.3.5 Análisis semántico

Esta etapa también agrega restricciones al conjunto de trayectorias de búsqueda del reconocimiento. Una forma en la cual las restricciones son utilizadas es vía un modelo dinámico del estado del reconocedor. Dependiendo del estado del reconocedor ciertas entradas sintácticamente correctas son eliminadas a consideración. Esto facilita realizar las tareas del reconocedor y eleva el desempeño del sistema.

## 3.4 Técnicas más utilizadas para el reconocimiento de voz

En la literatura científica asociada a esta área de conocimiento, se pueden encontrar diversas técnicas de clasificación de patrones de voz. Aquellas que mejores resultados han obtenido y las más prometedoras parecen que son las que a continuación se mencionan.

### 3.4.1 Comparación de plantillas o patrones

Este método consiste, en comparar el patrón a reconocer (de entrada) con una serie de plantillas o patrones que representan a las unidades a reconocer. La plantilla, no es más que un conjunto de características acústicas ordenadas en el tiempo (secuencia de vectores de parámetros o índices de una librería de centroides o *codebook*), y la comparación de patrones incluye un alineamiento temporal no lineal y una medida de distancia. Esta técnica, utilizada tanto para resolver problemas de reconocimiento de habla continua como aislada e incluso con una cierta independencia del locutor, se conoce como DTW (*Dynamic Time Warping*, por sus siglas en inglés).

El reconocimiento de patrones puede considerarse, de forma genérica, como una disciplina de la Inteligencia Artificial. Se analizan a continuación las razones de esta afirmación desde el punto de vista de la Inteligencia Artificial.

**Adquisición y representación del conocimiento.**

En Inteligencia Artificial, la adquisición del conocimiento consiste en la transferencia y transformación de conocimientos de una o más fuentes (libros, manuales, expertos, etc.) a un sistema informático y representarlo de forma útil para la máquina con vista a un tratamiento posterior. En reconocimiento de

formas se interpreta como la adquisición de patrones de clase conocida y su almacenamiento para establecer el patrón prototipo de cada clase.

**Aprendizaje.**

El aprendizaje implica cambios en el sistema que se adapta para permitir llevar a cabo la misma tarea a partir de las mismas condiciones de un modo más eficiente y eficaz cada vez. En un sistema de reconocimiento de formas, y dependiendo del método de aprendizaje se trata de calcular el patrón prototipo o el conjunto de patrones prototipo que caracterizan cada una de las clases a discriminar. Usualmente se utiliza un modelo de aprendizaje inductivo que se puede formular como sigue: una vez establecida la manera de representar el conocimiento y extraído éste, se calcula a partir de un conjunto de entrenamiento el patrón (o conjunto de patrones) prototipo utilizando un algoritmo de aprendizaje. Es necesario un esquema de evaluación que proporciona una medida de bondad del sistema.

**Clasificación.**

Consiste en proporcionar nuevos prototipos al sistema, independientes de los utilizados en el aprendizaje para que éste los etiquete utilizando el conjunto de clases disponibles.

**Evaluación**.

Toda clasificación lleva aparejada una medida de error, bondad o confianza. Deben proporcionarse mecanismos para evaluar esta bondad. Normalmente se utiliza un conjunto de patrones etiquetados por expertos y no usados en el aprendizaje.

El reconocimiento de patrones no es una técnica que se utilice muy a menudo para el reconocimiento de voz, pero ha sido una de las metodologías pioneras para que esta tecnología se vaya perfeccionando poco a poco.

### 3.4.2 Modelos ocultos de markov

Inicialmente, se comentará un poco de historia sobre los modelos ocultos de markov, se definen algunos conceptos importantes como: cadenas de Markov y modelos ocultos de Markov. Además, se muestra cuál es la relación que hay entre dichos modelos y los sistemas automáticos de reconocimiento de voz.

Desde finales de los 70´s, cuando los modelos ocultos de Markov fueron aplicados a sistemas de reconocimiento de voz, se desarrollaron técnicas para estimar probabilidades sobre estos sistemas en específico. Estas técnicas permiten a los modelos ocultos de Markov llegar a ser eficientes, robustos y flexibles de manera computacional. También les ha permitido servir de base de muchos sistemas de reconocimiento con vocabularios grandes e independientes del locutor por sus características alcanzadas. Los modelos ocultos de Markov fueron introducidos proponiendo este modelo como un método estadístico de estimación de las funciones probabilísticas de una cadena de Markov.

Esencialmente, los modelos ocultos de Markov son métodos para modelar sistemas con tiempo dependiente del comportamiento caracterizado por procesos comunes de corta duración y la transición entre ellos. Un modelo oculto de Markov puede ser pensado como una máquina de estado finito donde las transiciones entre los estados dependen de la ocurrencia de algún símbolo. Asociado con cada transición de estado una salida de la distribución de

probabilidad describe la probabilidad con la cual un símbolo ocurrirá durante la transición, y una probabilidad de transición indicando la probabilidad de esta transición.

En general, un modelo oculto de Markov es un subproceso estocástico asociado a una secuencia probabilística de símbolos. Este tipo de modelo está formado por una cadena de Markov y un conjunto de funciones de probabilidad asociada a cada estado; es decir, los estados ya no son simplemente símbolos sino que ahora se les asocian grupos con distribuciones de probabilidad.

#### 3.4.2.1 Definición

Los modelos de Markov describen un proceso de probabilidad el cual produce una secuencia de eventos o símbolos observables. Son llamados ocultos porque hay un proceso de probabilidad subyacente que no es observable, pero afecta la secuencia de eventos observados.

Los modelos ocultos de Markov pueden ser vistos como el modelo de un proceso, el cual produce una secuencia de eventos acústicos perteneciendo a una unidad específica, o palabra, en un vocabulario dado. Las variaciones entre las secuencias de observaciones de la misma clase, como la longitud de una palabra y pronunciación, son modelados por la naturaleza del elemento estocástico de un modelo oculto de Markov. Un sistema automático de reconocimiento de voz generalmente tendrá ***"N"*** modelos ocultos de Markov, uno para cada clase. El modelo reconoce la palabra cuando el estado final es alcanzado.

Entonces, se puede definir los estados de la cadena la cual genera datos observables donde se tiene:

- Un alfabeto de salida ***Y = {0,1,...,b-1}***.
- Un espacio de estados ***L = {1,2,...,c}*** con un único estado inicial ***$S_0$***.
- Una distribución de probabilidad de la transición entre los estados ***P (S' | S).***
- Una distribución de probabilidad de salida ***Q (Y | S,S')*** asociadas con las transiciones del estado ***S*** al estado ***S'.***

Entonces la probabilidad de observar una cadena de salida de un modelo oculto de Markov ***$Y_1$, $Y_2$,....,$Y_k$*** está dada por:

$$P(y_1, y_2, \cdots, y_t) = \sum_{s_0, s_1, \cdots, s_t} \prod_{i=1}^{t} P(s_i \mid s_{i-1}) q(y_i \mid s_{i-1}, s_i)$$

Ahora, se define lo que es un Modelo oculto de Markov de una manera más formal:

Sea $\lambda$ la quíntupla: $\lambda = \{N, M, \pi, A, B\}$

- donde ***N*** es el número de estados del modelo. y sea ***S = {1,2,3,...,N}*** el conjunto con ***N*** estados en el modelo.
- ***M*** es el número de símbolos ***(V)*** de observación y ***V = {$V_1$ ,$V_2$ ,...,$V_M$ }*** el conjunto con ***M*** símbolos.
- $\pi$ es la distribución de los estados $\pi = \{\pi_i\}$ y $\pi_i = P[q_1 = i], 1 \le i \le N$
- A es la probabilidad de transición de un estado a otro $A = \{a_{ij}\}$, $a_{ij} = P[q_{t+1} = j \mid q_t = i], 1 \le i, j \le N$ donde $q_t = i$ es el estado i en el tiempo ***t***

- B son las probabilidades de cada símbolo dado cada estado $B = \{b_j(k)\}$, $b_j(k) = P[O_t = v_k \mid q_t = j] \; 1 \le k \le M$ donde $O_t$ es la observación en el tiempo ***t.***

La notación compacta $\lambda = (A, B, \pi)$ es usada para representar a los modelos ocultos de Markov. Especificar un modelo oculto de Markov da la pauta para escoger el numero de estados, ***N***, tantos como números de símbolos discretos haya, M, y especificando el árbol de probabilidades de densidad, ***A, B,*** y $\pi$ .

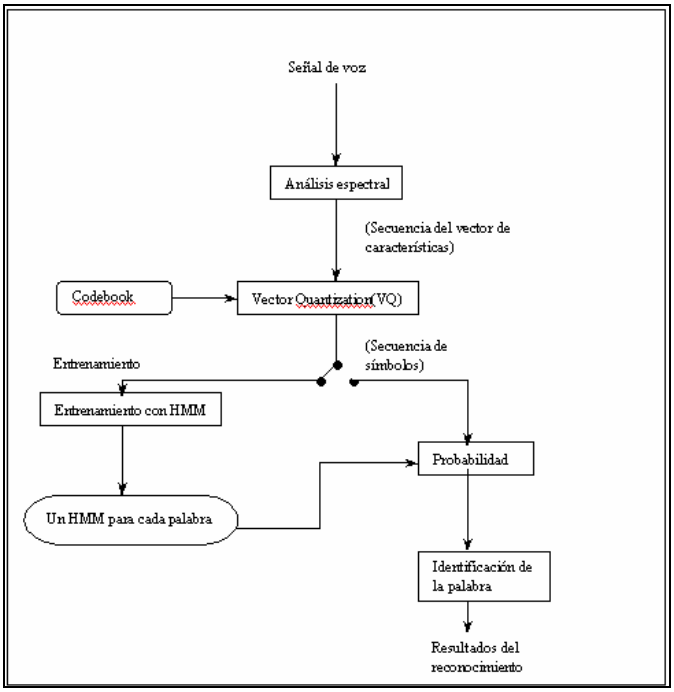
En la función de probabilidad existen dos diferentes funciones para calcular las probabilidades dependiendo del tipo de variable que se utilice, estas pueden ser discretas o continuas: en el caso de las variables discretas la función calcula una probabilidad por cada valor discreto de la variable, la función consiste en obtener distribuciones normales para los valores de las variables y haciendo uso de la media y de la varianza se puede calcular la probabilidad.

#### 3.4.2.2 Estructura general

La estructura general de un reconocedor basado en modelos ocultos de Markov puede ser vista como un módulo donde entra la señal de voz, normalmente ésta es almacenada como un archivo de sonido en formato wav. Después, pasa a un análisis espectral el cual produce una secuencia del vector de características, posteriormente se aplica la técnica de *Vector Quantization* (VQ) la cual se describe más adelante. En este paso se obtiene una secuencia de símbolos. En la fase de entrenamiento, se aplica este método y se crea un modelo para cada palabra o fonema según sea el caso. En la fase de

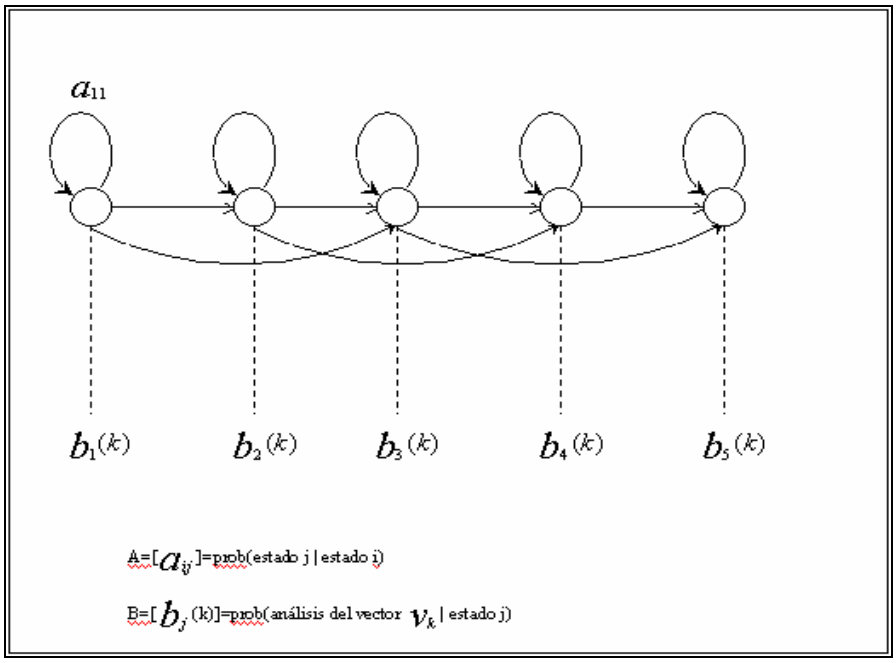
reconocimiento son calculadas las mayores probabilidades y se procede al reconocimiento de la palabra obteniendo los resultados de dicho reconocimiento. Este proceso se puede ver en la figura 4.

**Figura 4. Estructura general de reconocedor basado en modelos ocultos de markov**




En el enfoque de modelos ocultos de Markov, cada palabra es modelada por una red de transición la cual tiene un pequeño número de N estados, los cuales corresponden a las unidades fonéticas que se quieren reconocer (fonemas o palabras) existan. Cada estado físicamente corresponde en un sentido indeterminado a un conjunto de eventos temporales en la palabra pronunciada. En la figura 5 se muestra un ejemplo con 5 estados.

**Figura 5. Ejemplo de modelos ocultos de markov para reconocimiento de una palabra**




Todo el modelo de Markov es caracterizado por la probabilidad de la observación del símbolo $b_1(k)$ para el k-ésimo elemento del libro de código (codebook) en cada estado j , la probabilidad de transición del estado $a_{11}$ que describe como el nuevo estado j puede ser alcanzado desde el estado anterior i , y la probabilidad del estado inicial $\pi_1 \cdot \pi_1$ indica la probabilidad que el término i-ésimo estado, es el estado inicial de transición. $A=\{a_{11}\}$ y $B=\{b_1(k)\}$ son respectivamente llamados la matriz de transición de estados y matriz de frecuencia de probabilidad. A, B y $\pi=(\pi_1)$ consecuentemente representan variaciones temporales y espectrales en cada palabra. Todos ellos calculando y usando los datos de entrenamiento.

En la fase de reconocimiento para la salida desconocida, la probabilidad de la secuencia observada, es generada desde una secuencia de estados y calculada para cada palabra del vocabulario, y la palabra con la más alta probabilidad es seleccionada como la identificación correcta. El alineamiento en el tiempo es obtenido indirectamente a través de la secuencia de estados. No obstante aunque el proceso de entrenamiento es complicado, el proceso de reconocimiento es muy simple.

#### 3.4.2.3 Libro de código

En la codificación de la señal, cierto período es obtenido de la misma, y el patrón en este período es representado como un sólo código. Este procedimiento es realizado almacenando los patrones típicos de la señal (vectores de código o modelos), y asignando un código a cada patrón. La tabla que indica la correspondencia entre los patrones y los códigos es llamado libro de códigos. Donde una señal es comparada con cada patrón en un predeterminado intervalo, y la señal en cada período es delineada por un código indicando el patrón que es más similar a la señal.

Este libro también nos provee de un conjunto apropiado de patrones los cuales minimizan la distorsión cuando varios tipos de señales son representadas por un número limitado de patrones. Hay dos métodos de generación de libros de código, uno basado en aprendizaje aleatorio y otro basado en clusters. El primero permite a los vectores ser seleccionados aleatoriamente de los datos de entrenamiento y almacenados como vectores de códigos. El segundo método está basado en el algoritmo en que los datos de entrenamiento son agrupados en conjuntos no sobrepuestos y son calculadas las medias que minimizan la distorsión promedio. Estas medias son almacenadas como vectores de código.

#### 3.4.2.4 Vector quantization

Los cuantizadores de vectores operan sobre los vectores de características obtenidos de la señal de voz. VQ [3]; depende de la creación de un libro de código óptimo tal que la distorsión promedio remplazando cualquier vector de características por la entrada más cerca del libro de códigos es minimizada. Para un libro de código de tamaño Q indexado con q, el objetivo es seleccionar el conjunto de vectores del libro de códigos. La ventaja de VQ es que permite etapas subsecuentes en el reconocedor para que su complejidad disminuya. Estas etapas sólo necesitan direccionar las variaciones entre los vectores Q, o todos los que sean posibles. Las relaciones entre los vectores Q podrían ser pre-calculadas, salvando el tiempo de procesamiento. Estas relaciones pueden incluir métricas de distorsión, o la probabilidad que de que el vector $\vec{c}_q$ estará seguido por una entrada $\vec{c}_{qj}$ .

#### 3.4.2.5 Problemas del modelo

Existen tres puntos principales que deben ser resueltos cuando utilizamos los modelos ocultos de markov, los cuales se definen así:

- El primer factor a tomar en cuenta, es el problema de la evaluación del rendimiento del modelo. Para resolver este punto, se plantea la utilización de marcaje de cada palabra basada en la secuencia de observación para reconocer una palabra no conocida.
- Otra situación a considerar, es el problema de la secuencia de estados ocultos no cubiertos. La solución a este inconveniente es el desarrollo de un entendimiento del significado físico de los estados del modelo, para considerar los estados no cubiertos.

- El tercer escenario a considerar, es el problema del entrenamiento. Para esto, se emplea la obtención de forma óptima de los parámetros del modelo para cada palabra del modelo usando las pronunciaciones de entrenamiento.

Aunque no tienen ningún fundamento biológico, están bien adaptados al reconocimiento de grandes vocabularios independientemente del locutor. Sin embargo, el desempeño de dichos sistemas parece ser muy sensible al pre-procesamiento efectuado sobre la señal acústica de entrada. Aun así, los sistemas de reconocimiento con mejor desempeño en la actualidad se basan en los modelos ocultos de markov.

### 3.4.3 Redes neuronales

Las redes neuronales son estructuras de procesamiento paralelo de información, formadas por numerosos nodos simples conectados entre sí mediante pesos y agrupados en diferentes capas, entre las que se deben distinguir la capa de entrada y la capa de salida. Debido a su naturaleza intrínsecamente no lineal, a su capacidad de clasificación, y sobre todo a la capacidad que tienen para aprender una determinada tarea a partir de pares observación-objetivo sin hacer suposición alguna sobre el modelo subyacente, se han convertido en una de las herramientas más atractivas para la solución del problema del reconocimiento de habla.

Hoy en día se han conseguido resultados comparables a los obtenidos con otros métodos ya clásicos como los modelos ocultos de markov. Sin embargo, presentan diferentes problemas o inconvenientes como pueden ser: desconocimiento a priori de la estructura de capas y número de nodos necesarios para cada problema; un tiempo a veces excesivamente elevado

para su entrenamiento y la posibilidad de quedar "anclados" en mínimos locales de las funciones de coste usadas durante el entrenamiento de la red.

Además, la señal de habla requiere de métodos con capacidad de proceso en dos dimensiones, espacio y tiempo, y las redes neuronales, por sí solas, sólo tienen capacidad de procesado espacial. Esto obliga a combinar técnicas de programación dinámica así como modelos ocultos de markov con estas redes, consiguiendo modelar la variable tiempo, permitiendo no sólo la clasificaciones muy acertadas de las entradas de la red sino además la segmentación de la señal de entrada. Sin embargo, se han probado otras soluciones que incorporen a las redes algún tipo de memoria (finita, lazos de realimentación, o ambas), pero dificulta en gran medida el análisis de estas redes debido a su carácter no lineal.

#### 3.4.3.1 Fundamentos teóricos

**Neurona biológica**

Es la unidad morfológica funcional de realimentación del sistema nervioso, es una unidad procesadora de información, que funciona de la siguiente manera: recibe información del entorno (o de otras neuronas), realiza algún procesamiento y emite una respuesta. Se estima que en el cerebro humano hay más de cien mil millones (1011) de neuronas y en un área de un milímetro cuadrado se encuentran unas cincuenta mil. La neurona biológica tiene variedad de formas y tamaños, esta está compuesta por:

**Cuerpo Celular**

Contiene el núcleo y es el encargado de realizar las actividades metabólicas de las neuronas.

**Dendritas**

Son estructuras que parten del cuerpo celular. Se especializan en la recepción de señales (impulsos eléctricos) de otras células nerviosas.

**Axón**

Se encarga de la propagación de impulsos a otras células nerviosas, o sea, lleva la salida de la neurona a las dendritas de otras neuronas.

**Sinapsis**

En realidad no es una parte de la neurona, sino más bien el proceso de conmutación que ocurre entre ellas, este proceso puede tener un efecto excitatorio o inhibitorio. En la sinapsis ocurre la transferencia de impulsos eléctricos desde el axón de una neurona a la dendrita de la otra, debido a la liberación de un neurotransmisor (iones).

Estudios sobre la anatomía del cerebro humano concluyen que hay más de mil sinapsis a la entrada y a la salida de cada neurona. Es importante anotar que aunque el tiempo de comunicación de la neurona (unos pocos milisegundos) es casi un millón de veces menor que en los actuales elementos de los computadores, ellas tienen una conectividad de miles de veces superior que los actuales supercomputadores. Las neuronas y las conexiones entre ellas (sinapsis) constituyen la clave para el procesamiento de la información.

**Características**

El cerebro humano tiene muchas características deseables en un sistema artificial, entre ellas se pueden mencionar:

- Tolerancia a fallas, diariamente mueren neuronas sin que esto afecte el desempeño general del sistema.
- Flexibles, se ajustan a nuevos ambientes por aprendizaje.
- Maneja información ambigua, múltiple e inconsistente.
- Altamente paralelo.
- Robusto, compacto, seguro y consume poca energía.
- Alta velocidad de respuesta.

#### 3.4.3.2 Red neuronal artificial

Desde el punto de vista funcional, una red neuronal es un sistema de procesamiento de información físico o algorítmico, formado por un gran número de elementos computacionales muy simples (nodos neuronales), cada uno propietario de una cantidad de memoria local y conectada a través de canales de comunicación unidireccionales (uniones o arcos). Estos últimos son ponderados, es decir, algunas uniones entre nodos son más fuertes que otras, y transportan únicamente información numérica.

Las redes neuronales se denominan también sistemas de procesamiento paralelo distribuido o memorias asociativas debido a que llegan a resultados "inteligentes" realizando muchos cálculos paralelos independientes y sin seguir reglas lógicas rígidas. La respuesta la produce el circuito activo o función de transferencia que forma parte del cuerpo de la neurona. Las "dendritas" llevan las señales eléctricas al "cuerpo " de la misma. Estas señales provienen de sensores o son las salidas de neuronas vecinas. Las señales por las "dendritas" pueden ser un voltaje positivo o negativo, los positivos excitan el cuerpo de la neurona y por lo tanto la generación de una señal de salida. Los voltajes negativos inhiben la respuesta de la neurona.

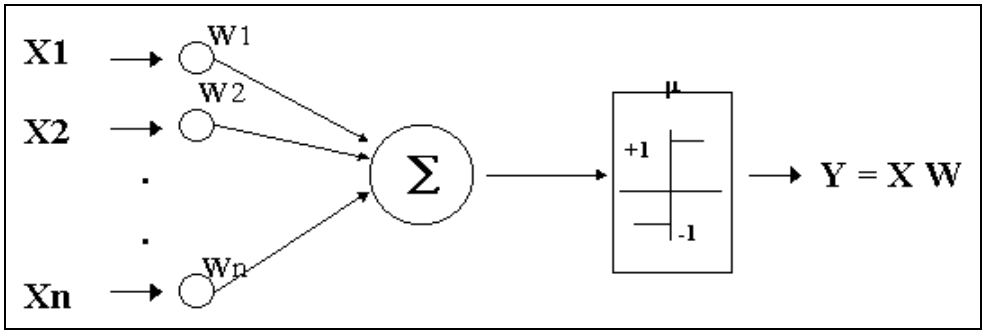
De esta manera la representación de una neurona artificial es un circuito eléctrico que realiza la suma ponderada de las diferentes señales que recibe de otras unidades iguales y produce en la salida un uno (1) o un cero (0) según el resultado de la suma con relación al umbral o nivel de disparo. Entre las más importantes se encuentran:

- Son capaces de manejar entradas con ruido, inexactas, probabilísticas y difusas (filtran la información que manejan).
- Responde de manera simultánea a diversas situaciones en forma paralela.
- La potencia de las conexiones entre neuronas es usada para representar conocimiento.
- Se entrenan, auto-organizan, aprenden y olvidan.
- Son robustas y tolerantes a fallas, la falta de una o varias neuronas no implica un fallo total en la red Neuronal.
- Son flexibles, lo que les permiten adaptarse fácilmente a nuevos ambientes, ya que pueden catalogarse como sistemas inteligentes.
- La velocidad de respuesta es menor en comparación con la del cerebro humano.

#### 3.4.3.3 Fundamentos básicos para funcionamiento

Existe un bloque de construcción fundamental para todas las redes neuronales. El modo general de operación es como se muestra en la figura 6.

**Figura 6. Neurona Artificial individual**




La suma ponderada de las entradas es la salida de la neurona, la cual está expresada como: Y= ø{(W1.X1 - μ) + (X2 . W2 - μ) + . . . + (Xn . Wn - μ)}, y de forma simplificada como:

$$Y = \theta\left(\sum_{j=1}^{n} X_j.W_j - \mu\right)$$ (1), donde X1, X2, . . ., Xn son el conjunto de entradas aplicadas a través de un conjunto de pesos asociados W1, W2, ...,Wn a la neurona. Estas entradas corresponden a los niveles de estimulación y los pesos a las intensidades sinápticas de la neurona biológica, además, n es el número de entradas, μ es el umbral. La función de activación:

ø (a) = +1, si a >= 0

-1, si a < 0 , donde: $\sum_{j=1}^{n} X_j.W_j - \mu$

En otras palabras, el modelo realiza una suma ponderada de las entradas. Si la suma es mayor o igual que el umbral, entonces la neurona se dispara y la salida es +1, en caso contrario es -1. La función ø(a) se denomina función de

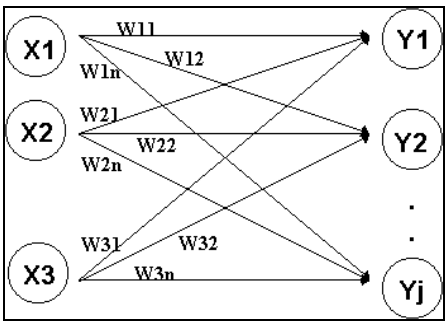
activación, en este caso es un limitador duro (sólo toma dos valores). El umbral μ se puede usar como un peso adicional Wo conectado a una entrada con valor constante Xo = +1. La expresión (1) se reduce a:

$$Y = \theta\left(\sum_{j=1}^{n} X_j.W_j\right)$$, Donde Wo = - μ y Xo = +1

#### 3.4.3.4 Red neuronal multicapa

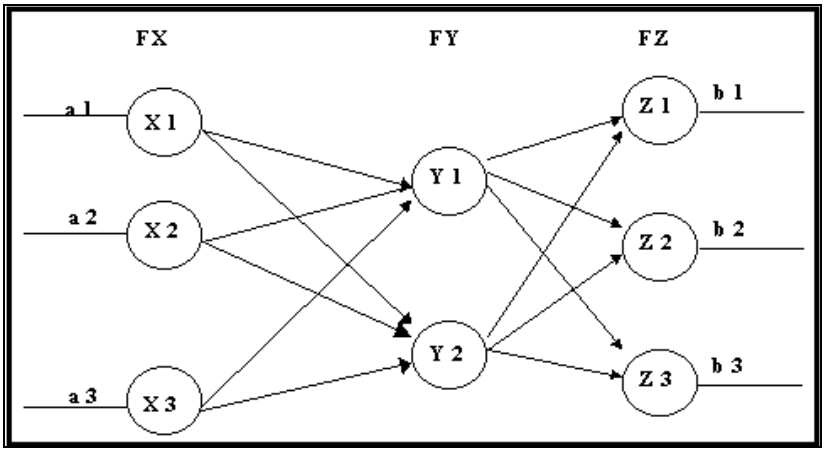
Las neuronas artificiales pueden ser combinadas como en la red de la figura 7.

**Figura 7. Combinación de entradas y salidas, red neuronal**




Todas las entradas están totalmente conectadas al conjunto de neuronas, es decir, Wij parte desde cada entrada Xi a cada neurona Yj. En este caso, las salidas son computadas así: $Y_j = \sum W_{ij}.X_i$ o en notación vectorial como: Y = X. W. Este tipo de arquitectura ha permitido la construcción de redes en cascada, formadas por dos o más capas de neuronas. La operación de estas redes se duplica con relación a las redes monocapa. En la figura 8 que se muestra a continuación se muestra un ejemplo de una red neuronal de tres capas.

**Figura 8. Red neuronal multicapa**




Estas redes son llamadas Fx, Fy y Fz. La capa Fx es igual a X1, X2 y X3, y se denomina capa de entrada; la capa oculta Fy (Y1 y Y2), y la capa de salida Fz (Z1, Z2 y Z3). Las conexiones entre nodos se representan mediante flechas dirigidas. Obsérvese que en este caso, existe una conexión desde cada nodo Fx hasta cada nodo Fy, así como desde cada nodo Fy hasta cada nodo Fz, a este tipo de conexión se le denomina redes *Feed Forward* o de propagación hacia delante. Pueden presentarse posibles conexiones intercapa, es decir, entre nodos de una misma capa. La disposición de los nodos y las conexiones de una red neuronal constituyen lo que se denomina topología.

Cada nodo de una red neuronal recoge los valores de todas sus conexiones de entrada, realiza una operación matemática predefinida, generalmente una combinación lineal seguida de una operación no lineal de activación, y produce un valor de salida único e independiente. Así mismo, cada conexión entre nodos tiene su propio valor o peso, establecido durante el proceso de entrenamiento, definido desde un comienzo o asignado en forma aleatoria. Estos pesos almacenados, son los que le permiten a la red neuronal "aprender" y "recordar". Las conexiones de valor cero no tienen efecto.

En adición a los elementos de procesamiento y a las conexiones ponderadas entre ellos, otros elementos básicos en el diseño, implementación y utilización de las redes neuronales son las funciones de activación y los datos o patrones de entrada y salida, que son respectivamente (a1, a2, a3) y (b1, b2, b3) mostrados en la Figura 8. Una vez propagados, los patrones de entrada son procesados por la red y ésta produce como resultado unos patrones de salida representativos, por ejemplo la secuencia apropiada de comandos que producen la respuesta deseada de un controlador o el conjunto de variables que representan la solución a un problema de optimización específica.

Las funciones del umbral o de activación introducen un cierto grado de alineabilidad en la dinámica de la red y limitan sus valores de salida dentro de un rango específico. Aunque son posibles muchas funciones de activación para redes neuronales, las más utilizadas son la sigmoide y la gaussiana, muy comunes en estadística. Otras opciones incluyen la función lineal, la función escalón y la función rampa. Sin la alineabilidad introducida por la función de activación, las redes neuronales no serían tan potentes ni tendrían la capacidad de generalización que las caracteriza.

En otras palabras, las redes neuronales aprenden de ejemplos, de manera similar como un niño reconoce un perro en la calle a partir de un ejemplo de perro visto en la televisión o en un libro. Esta capacidad estructural de generalización es uno de los aspectos más interesantes de las redes neuronales. Una vez que una red neuronal ha sido entrenada con ejemplos apropiados, puede ser utilizada como modelo para predecir las variables de salida deseadas de un proceso, tomar decisiones a partir de una gran cantidad de información de entrada disponible, encontrar la solución óptima a un problema de control, reconocer una palabra escrita, obedecer a un comando

hablado, rechazar productos defectuosos en una línea de alta producción y realizar otras tareas concretas.

#### 3.4.3.5 Tipos de aprendizaje

Existen dos métodos de aprendizaje aplicables a las redes neuronales, ellos son:

##### 3.4.3.5.1 Supervisado

Las redes de entrenamiento supervisado han sido los modelos de redes más desarrollados desde el inicio de estos diseños. Estas necesitan de un profesor humano que las oriente, le informe de sus errores, estimule sus logros y decida el momento de concluir el entrenamiento o la intensidad y frecuencia de presentación de los ejemplos.

Uno de los algoritmos más populares de entrenamiento supervisado es el “Backpropagation” y su funcionamiento está dado por:

- Presentación de entradas a la red.
- Producción de una salida por parte de la red.
- Comparación de la salida de la red con la respuesta deseada, estableciendo una medida de error con la cual se modificarán los pesos de las conexiones sinápticas de la red.

Este proceso se hace con el fin de que al presentar de nuevo los datos a la red, su respuesta mejore aproximándose a la respuesta correcta. El aprendizaje supervisado es adecuado para tareas de aproximación, heteroasociación y control.

#### 3.4.3.5.2 No supervisado

Son autónomas, trabajan solamente con información local. El conjunto de datos de entrenamiento consiste sólo en los patrones de entrada. La red aprende a adaptarse basada en experiencias recogidas de los patrones de entrenamiento anteriores  Este tipo de entrenamiento es adecuado para redes de clasificación y auto-asociación.

### 3.4.3.6 Aplicaciones

Las redes neuronales pueden realizar cualquier función que normalmente se asigna a un computador. Son adecuadas para reconocer, clasificar, completar, optimizar y realizar funciones complejas de manipulación de patrones; otras aplicaciones incluyen la toma de decisiones a partir de datos masivos, el modelo y mapeo de sistemas no lineales, y en general dan solución a problemas no numéricos. En la actualidad se dispone de procesadores neuronales como el *Ni1000 de Néstor* e Intel®, los cuales son capaces de realizar varios billones de operaciones por segundo, reconociendo miles de patrones. Algunos de los principales campos de operación de estos procesadores son: el comercio en la predicción de ventas, las finanzas para el reconocimiento de firmas, en la medicina para el procesamiento de imágenes y señales, el reconocimiento de caracteres utilizado en el fax y el correo, y en la seguridad para el reconocimiento de huellas dactilares.

Las Redes Neuronales Artificiales tienen su aplicación en algunas áreas del saber como la biología, para descripción de sistemas microbianos e interpretar secuencias de nucleótidos, en la física se pueden utilizar para modelar los fenómenos de la termodinámica y la mecánica estadística, y en la ingeniería en las disciplinas de la microelectrónica, robótica, comunicaciones,

electromedicina, entre otras. Cada vez será tema de más interés para los filósofos, sociólogos, políticos, economistas; que de alguna forma se verán beneficiados con los avances y desarrollos que tengan las Redes Neuronales Artificiales.

Estas metodologías anteriormente explicadas, han incursionado de manera efectiva en ésta materia, con aplicaciones como sistemas de restauración y generación del habla, complejos sistemas de reconocimiento de palabras aisladas y de texto continuo. Por esta razón muchos de los productos que existen actualmente en el mercado han sido resultado de la investigación, desarrollo y perfeccionamiento de éstas técnicas aplicadas al procesamiento de voz. Para finalizar, en la actualidad el reconocimiento de voz es un campo de investigación con objetivos, métodos y aplicaciones bien definidos, en el que hay mucho trabajo por realizar en distintos niveles (teórico, práctico, etc.) y en distintas materias como: procesamiento de señales, acústica, fonética, reconocimiento de formas e Inteligencia Artificial.

# 4. INTEGRACIÓN DE LA VOZ EN INTERFACES GRÁFICAS MULTIMODALES

El tópico de expresar ideas por parte de los seres humanos es muy extenso y se encuentran muchas formas de expresión, por ejemplo, el lenguaje hablado, el lenguaje escrito, los gestos, entre otras. Estos elementos forman una interfaz entre las personas muy importante, interfaz que se ha llevado a nivel hombre-máquina. Sin embargo, éstas formas de expresión, tienen sus ventajas y desventajas, la cuales también se transfieren a la interfaz con la máquina.

Entre las formas de comunicación más comunes que existen, se encuentra la voz. La voz, es una forma natural de comunicación que es persuasiva, eficiente y puede ser usada a distancia. Este tipo de comunicación, es una manera rápida de expresar ideas para cualquier hablante, debido a que se aprende desde el mismo entorno.

Los beneficios de interfaces de usuario manejadas por voz han sido apoyados ya desde hace varios años. Debido a esto, se han orientado varios esfuerzos en investigar el comportamiento de la voz como un dispositivo de entrada auxiliar en interfaces hombre-máquina. A pesar de eso, la total aceptación de las interfaces hombre-maquina por medio de la voz, es un hecho aun por ocurrir, debido a las desventajas que presenta esta modalidad, el lenguaje humano y la forma de reconocimiento de estos comportamientos por parte de la máquina.

No obstante, se ha planteado un modelo complementario de comportamiento, sugiriendo que las interfaces hombre-máquina de reconocimiento de voz se vean complementadas con otra u otras modalidades, la cual permita suplir en parte las desventajas que el reconocimiento de voz pueda presentar. Con ello, las modalidades que pueden ser involucradas en un interfaz hombre-maquina de reconocimiento de voz, tendrán recíprocas fortalezas y debilidades las cuales pueden ser incrementadas en una interfaz multimodal de usuario. Combinando dos o más modalidades, las fortalezas de una pueden ser usadas para compensar las debilidades de las otras.

Este modelo complementario de comportamiento, involucra varias ideas, términos y conceptos, los cuales, es necesario aclarar para tener un mejor panorama de lo que se plantea y se pretende con él.

## 4.1 Naturaleza de las investigaciones multimodales

En el sentido más básico, se tiene conocimiento que multimodalidad es combinar por lo menos dos modalidades para entrada de datos, teniendo en cuenta que cada uno requiere un conocimiento intensivo y diferentes componentes tecnológicos. A menudo, esto ha significado combinar la modalidad de voz con tecnologías basadas en la vista o entrada de pluma electrónica, desde que los sistemas de voz/pluma y voz/labios multimodales, son los que actualmente ocupen la atención dentro de este campo.

En ambos tipos de sistemas, la meta explícita ha sido integrar modalidades complementarias de manera que se produzca una mezcla sinérgica, tal que cada modo sea capaz de acoplarse y a la vez, superar debilidades del otro modo. El aprovechamiento del diseño de estos sistemas, ha sido apoyado por la filosofía de utilizar modos y componentes tecnológicos

para explotar la ventaja del lenguaje natural, y de combinarlos de alguna manera que logren una compensación mutua. Una ventaja de lograr tal mezcla es que la arquitectura multimodal resultante puede funcionar más robustamente para las tecnologías basadas en reconocimiento individual, tal como sistemas de reconocimiento de voz.

Otra característica que se encuentra dentro de las investigaciones multimodales y que es casi obvia, es la multidisciplina. Dentro de las innovaciones de sistemas multimodales claramente requiere una especialización multidisciplinaria en diferentes áreas más allá de la informática, incluso la ciencia de percibir y escuchar, la percepción y la visión, lingüística, psicología, procesamiento de signos, reconocimiento de patrones y estadísticas.

La naturaleza multidisciplinaria que se encuentra en los sistemas multimodales, conlleva varias implicaciones. Evolucionar con éxito a este campo, significa que la informática necesitará ponerse en una panorámica más abierta desde su punto de vista global.

Como resultado, se tiene un concepto de “comunidad constructora” entre investigadores multimodales y se necesita que se tengan las relaciones necesarias entre aquéllos que representen las claves tecnológicas de los componentes, las disciplinas académicas, y las diferentes perspectivas en el ambiente cultural e internacional.

Entre las metas mas importantes de la investigación de la multimodalidad se ha visto incrementado la accesibilidad de interfaz para usuarios diversos y usuarios no especializados que eficazmente trabajen hacia borrar las fronteras entre los usuarios instruidos y los no-instruidos en ese sentido. Las interfaces

multimodales aumentan la accesibilidad de informática por los usuarios de edades diferentes, nivela la habilidad, nivela los impedimentos sensores y motores, idiomas nativos e incluso enfermedades temporales.

Hay grandes diferencias individuales en habilidad y preferencias de la personas para usar modos diferentes de comunicación, y una interfaz multimodal le proporciona opciones de la interacción al usuario, lo cual oculta las limitaciones personales del mismo.

La investigación de la multimodalidad representa una ciencia sin fronteras ya que requiere de un amplio conocimiento para combinar:

- dos o más modalidades que representan diferentes componentes tecnológicos
- las diversas perspectivas multidisciplinarias, y
- una perspectiva multicultural/internacional natural en modelos de comunicación.

Además, la investigación multimodal está borrando fronteras y con ello promueve la accesibilidad de la informática para:

- usuarios diversos y no especializados, y
- los contextos de uso variados

Se debe tener una perspectiva de lo que se puede lograr a través de la buena disposición para cruzar fronteras, intelectualmente y literalmente. Debido grandemente a la sed para cruzar fronteras, las investigaciones multimodales han tenido un consolidado avance con velocidad notable durante los últimos años. Sin embargo, manteniendo esta proporción de progreso, es una

advertencia de que continuará requiriendo trabajo en equipo multidisciplinario, la unificación de diferentes habilidades, perspectivas, destreza, se plantea en una visión a largo plazo y la multimodalidad como una "comunidad constructora". Como el planteamiento y la aplicación de nuevos tipos de sistemas multimodales están en aumento, la sofisticación de éstos se ve con un aumento considerable, pero la meta será un cambio en el equilibrio de interacción de la interfaz humano-computadora muy íntimo al humano.

## 4.2 Diferencia entre sistemas multimodales y multimedia

Los sistemas multimodales y multimedia, a pesar de la similitud de nombre, tienen diferencias sustanciales, las cuales se deben considerar para entender a profundidad cada uno de los términos. Debido a los diferentes usos de terminología en nuestro lenguaje, las expresiones modalidad, medio y modo, se definen por separado y se debe tener claro cada uno de estos conceptos. Adicionalmente, se debe introducir el término de multimodalidad, el cual se basa en el uso de modalidades por medio de las cuales los humanos reciben o envían información.

Estas modalidades pueden ser táctil, visual y auditiva, lo cual, puede involucrar el uso de por lo menos dos modalidades para presentar la información, por ejemplo, la actividad verbal o manual. Así, en una interacción multimodal un usuario puede recibir información por medio de la vista y el sonido, y responder por medio de la voz y el tacto. De igual forma, se puede comparar multimodalidad con unimodalidad.

La unimodalidad, se basa en el uso de una sola modalidad para recibir o presentar información. Para ejemplificar el concepto anterior, se puede decir que en una presentación multimedia, el usuario recibe una presentación para la

vista y responde por medio del teclado. Estos sistemas que responden a la característica de unimodalidad, se les llama sistemas interactivos, representaciones múltiples o sistemas multimedia. Dichos sistemas, comparten un objetivo común con los sistemas multimodales: la interacción eficaz con el usuario.

Para que un sistema se le considere que cuenta con una interacción eficaz con el usuario, no debe ser únicamente fácil de usar, debe también apoyar al usuario en la ejecución de la tarea. Independientemente de las diferencias tecnológicas en la aplicación, estos sistemas apuntan siempre al apoyo a usuarios mientras ejecutan tareas en particular.

Sin embargo, sistemas multimedia y multimodales tienen diferencias importantes. Los sistemas multimedia tratan sobre la presentación de información. En cambio, los sistemas multimodales interpretan y generan información para ser presentada en diferentes modalidades. Visto desde un enfoque de usuario, la distinción entre una interfaz multimedia y multimodal está basada en las capacidades que tiene cada sistema para sus entradas y salidas. De ésta forma, una interfaz multimodal proporciona al usuario múltiples entradas y salidas en su sistema, por ejemplo, puede usar el lenguaje acompañado de una acción de una pluma electrónica.

Una interfaz multimedia soporta únicamente múltiples salidas en su sistema, por ejemplo, un usuario puede recibir una salida de texto y audio a una entrada proporcionada por medio del tacto. De lo anterior, se puede concluir que los sistemas multimedia son un subconjunto de los sistemas multimodales. Una diferencia alternativa entre sistemas multimedia y multimodales puede ser basada en la experiencia de la interacción. Visto desde un punto de vista de sistemas, un sistema multimedia es también multimodal debido a que provee

diferentes medios al usuario una salida multimodal, por ejemplo, información visual y audible, y entradas multimodal, como por ejemplo, escribir por el teclado o dar un clic al ratón. Sin embargo, estos sistemas no pueden responder a una combinación, por ejemplo, teclado y ratón, ya que no son adaptables a diferentes usuarios. Además, mientras se interactúa con un sistema multimodal, el usuario recibe entradas multimodales y es capaz de responder por las mismas modalidades porque tiene un conocimiento de la interacción. Mientras en los sistemas multimedia un usuario se adapta a las capacidades de percepción, los sistemas multimodales se adaptan a las necesidades y preferencias del usuario.

Este argumento, sin embargo, apunta a destacar la importancia de la experiencia interactiva y no la importancia del usuario. Si se basa sólo en el usuario, un sistema podría ser multimodal para un usuario y multimedia para otro. Por otra parte, en las investigaciones sobre sistemas multimodales, a menudo se asume esa comunicación del humano-humano como 'maximización multimodal y multimedia'. El 'valor agregado' de los sistemas multimodales, es debido a que falta investigación sobre porqué necesitamos desarrollarlos. Las interfaces multimodales se basan en la naturalidad de comunicación entre el usuario y el sistema.

Naturalidad se refiere a que una comunicación humano-maquina debe ser como una comunicación humano-humano. Es por ello, el enfoque en la realización tecnológica por técnicas del reconocimiento generadoras de lenguaje natural, gestos, táctiles, etc. El objetivo principal, es proporcionar a los usuarios un sistema que sea capaz de emular cómo los humanos actúan recíprocamente el uno con el otro. Se debe explotar la capacidad que tenga la maquina de percibir la información que el humano le provea, a manera de que pueda aprender y luego pueda presentar información con significado. No

obstante, hay diferencias entre la interacción humano-humano y la interacción humano-computadora. En la interacción humano-humano, por ejemplo, hay disponible un sistema bastante sofisticado llamado mente humana, que indica que modalidad usar y cuándo. Las actuales investigaciones sobre sistemas multimodales, a menudo asumen modalidades para esta tecnología, mientras se ejecutan tareas en particular sin preguntar porqué.

En resumen, se ha enfocado las diferencias entre multimodalidad y sistemas multimedia. Los sistemas multimedia se refieren a la adaptación de las capacidades preceptúales del sistema respecto a los usuarios. Los sistemas multimodales son los que apoyan con diferentes modalidades de entrada y salida a los usuarios, según las preferencia y necesidades del mismo.

Por otra parte, para cualquier sistema es importante su etapa de diseño y modelación, ya que de esto depende en buena parte la funcionalidad del sistema, por lo cual, a continuación se presenta este tema como continuidad de del planteamiento inicial propuesto en este capítulo.

## 4.3 Modelación y diseño de los sistemas multimodales

Durante una comunicación multimodal natural, el ser humano habla, gesticula, mira y se instala un flujo poderoso de comunicación que para nada es igual a una interfaz de usuario gráfica con un teclado o unos clics de un ratón. Un profundo cambio está ocurriendo para llevar a los usuarios a una conducta natural siendo el centro un interfaz hombre-máquina.

La realización de interfaces multimodales debe permitir controlar favorablemente conductas comunicativas a través de interacciones con el

sistema, a modo de que esta experiencia sea transparente para el usuario, como si se tratara de la vida misma.

La voz, las manos y el cuerpo entero juntos, una vez aumentado por sensores como micrófonos y cámaras, son de los últimos dispositivos de entrada multimodal transparentes y móviles.

El área de sistemas multimodales se ha extendido rápidamente durante los últimos años. Subsecuentemente, desde que surgió la idea “Put that there” (Ponga esto allí) que procesaba voz y manipulación de objetos, se ha avanzado con pasos largos desarrollando más sistemas multimodales generales. Como un punto de partida para avanzar en nuevos sistemas multimodales, el trabajo proactivo empírico ha generado información predictiva de la interacción multimodal de humano-máquina, que se utiliza para guiar planes de sistemas multimodales.

El progreso mayor ha ocurrido en el hardware y software para las tecnologías de los componentes como el habla, escritura y visión. Además, se han establecido componentes arquitectónicos y áreas de trabajo más generales para diseñar más sistemas multimodales. En este sentido, la industria se enfrenta con nuevos retos. Los sistemas que dominan, por ejemplo, equipos de audio, se reemplazarán en un futuro cercano por multi-sistemas que se enfoquen en un dominio en los que se integren medios de comunicación diferentes. Además, se desarrollan estilos de interacción multimodal para operar estos sistemas fácilmente.

Se debe dirigir estos retos en la reutilización del desarrollo de dispositivos o componentes de estos sistemas. La noción de dispositivos o componentes, debe ser interpretada en un sentido amplio e incluye especificaciones de

software y componentes determinados, modelación de la tarea, y guías específicas.

Es importante que esta tecnología, se adapte en un interfaz hombre-maquina de forma intuitiva y natural, como la que se maneja en una interacción humano-humano. Esta interacción incluye, que estén bien establecidos los dispositivos de entrada de la máquina para cada una de las modalidades que se incluyan. Así como en una comunicación humano-humano, estas modalidades deben usarse para la transmisión de información y, al mismo tiempo, transmitir un “estado de usuario” del compañero de comunicación durante la interacción.

La habilidad de desarrollar sistemas multimodales robustos, dependerá del conocimiento de los modelos de integración naturales, que representan la combinación de modalidades de uso de las personas. Se debe tener un punto de partida teórico y de diseño. Con ello, se intenta tener un desarrollo de modelos predictivos de la integración de modalidades naturales para dar lineamientos a los planes sobre arquitecturas multimodales.

### 4.3.1 La ciencia cognoscitiva en el diseño

La habilidad de desarrollar sistemas multimodales, depende del conocimiento de la integración natural de modelos, que representen combinaciones de modalidades de entrada diferentes que son utilizadas por las personas.

En particular, el diseño de los nuevos sistemas multimodales, dependen del conocimiento íntimo de las propiedades de los diferentes modos, el volumen de información que ellos llevan, las características particulares del idioma multimodal, su forma de procesamiento y la integración y características de

sincronización para los usuarios para la interacción. También se debe tomar en cuenta la predicción exacta, cuándo es probable que los usuarios actúen recíprocamente de forma multimodal, y cómo usuarios diferentes utilizan modelos iguales de integración específicos.

La relevancia de la literatura cognoscitiva en estos temas es muy extensa, sobretodo cuando se da consideración, a todas las percepciones de los sensores y capacidades, que involucran modos de entrada diferentes a los que actualmente se está incorporado en nuevas interfaces multimodales. De esta forma, se introducirán los temas principales sobre la ciencia cognoscitiva y datos encontrados, que son importantes resaltar dentro de los sistemas multimodales más comunes.

**Reciprocidad de usuarios multimodales**

Durante la comunicación interpersonal natural, las personas están actuando recíprocamente siempre multimodalmente. Por supuesto, en este caso el número de fuentes de información o modalidades que un interlocutor tiene disponible para supervisar es esencialmente ilimitado. Sin embargo, todos los sistemas multimodales están restringidos en el número y tipo de modalidades de entrada que pueden reconocer.

También, un usuario puede crear una entrada activa durante la interacción hombre-maquina que puede ser multimodal o usando simplemente una modalidad. Es decir, aunque los usuarios pueden tener una preferencia fuerte en general a actuar recíprocamente multimodalmente y no unimodalmente, no es ninguna garantía que ellos emitirán cada orden en la multimodalidad del sistema, dependiendo del tipo de interfaz disponible.

Por consiguiente, la primera pregunta no trivial que se realiza durante el procesamiento del sistema, es si un usuario se comunica unimodalmente o multimodalmente. En particular, los usuarios casi siempre se expresan multimodalmente al describir información espacial sobre alguna situación, numerando, clasificando según tamaño, dando una orientación o forma de un objeto.

Se debe dar énfasis, a que los sistemas multimodales del futuro necesitarán distinguir entre los casos cuando los usuarios utilicen o no la multimodalidad para comunicarse, para que las decisiones exactas puedan tomarse sobre entradas paralelas, que deben interpretarse conjuntamente versus individualmente.

Asimismo ese conocimiento del tipo de acciones debe ser incluido en una aplicación, ya que si la aplicación trae consigo manipulación de información especial, debe influir en la decisión básica para diseñar y construir un interfaz multimodal absoluta.

En una interfaz multimodal, que procesa modalidades de entrada pasivos o mezclados, hay por lo menos siempre uno pasivamente-activo que es la fuente de entrada que proporciona información continua, por ejemplo, seguimiento con la mirada, posición de la cabeza. En estos casos, toda la entrada del usuario tiene que ser por definición multimodal, y el problema primario se volvería la segmentación e interpretación de cada flujo de entrada continuo en las acciones que son de importancia para la aplicación.

En el caso de interfaces multimodales combinadas, por ejemplo, mirada que rastrea y entrada con un clic del ratón, aún puede ser oportuno distinguir las formas de entrada de usuario activas con lo cual, se puede considerar en

base a la precisión y eficacia el manejo de los eventos, incluso, llegando por estas mismas condiciones a tomarlos como unimodales.

**Integración y características de la sincronización de entradas multimodales de usuario**

Hasta el momento, los sistemas multimodales se han enfocado principalmente en la selección simple de objetos o situaciones en un despliegue, en lugar de considerar el tema sobre una visión más amplia como lo es la integración de modelos multimodales.

Desde del sistema "ponga esto allí", hablar y apuntar ha sido la forma prototípica de integración multimodal. Se basa en un proceso semántico de una entrada hablada, pero el significado de un término "que", era resuelto procesando la coordenada del x,y indicada, apuntando a un objeto. Desde ese tiempo, otros sistemas multimodales también han intentado resolver expresiones que usan un acercamiento similar, por ejemplo, usando una situación de ver en lugar de apuntar manualmente.

Desgraciadamente, este concepto de interacción multimodal visto como “sólo apunte y hable” ha limitado el uso de nuevos modos de entrada para la selección de objetos, así como se hace con el ratón. Esto representa la persistencia de una metáfora orientada al ratón, que ya tiene años de vigencia.

En contraste, modalidades que transmiten entradas escritas, manuales en gesticulaciones y las expresiones faciales, son capaces de generar información simbólica que es mucho más expresiva que una simple acción de apuntar o seleccionar. Juntos, la ciencia cognoscitiva y los datos modelados por usuario resaltan el hecho que cualquier sistema multimodal diseñado para procesar

“hablar y apuntar” exclusivamente, no les proporciona funcionalidad muy útil a los usuarios.

Por esta razón, los algoritmos especializados en procesar este tipo de relaciones, se han limitado en el uso práctico del diseño de sistemas multimodales futuros. Está claro que un campo más amplio de la integración multimodal, necesita emitirse y dirigir trabajo futuro. La investigación futura también debe explorar modelos de integración típicos entre otras combinaciones de modalidades viendo hacia el futuro, como el habla y la vista.

Para abreviar, aunque dos modos de la entrada pueden ser muy interdependientes y pueden sincronizarse durante una interacción multimodal, la sincronización no implica simultaneidad. La evidencia empírica revela, que los signos multimodales no ocurren simultáneamente en la mayoría de las veces durante la comunicación hombre-computadora o propiamente en la comunicación humana. Por consiguiente, diseñadores de sistemas multimodales en un futuro, no necesitan contar con signos convenientemente solapados para lograr un proceso exitoso, en la construcción de las arquitecturas que se necesite.

En el diseño de nuevas arquitecturas multimodales, es importante hacer notar los datos en el orden que se han usado las modalidades de entrada y retrasos de tiempo promedio, para determinar la probabilidad que una expresión es multimodal versus unimodal, y establecer términos de referencia temporales para la fusión de entradas. En el diseño a futuro, las arquitecturas multimodales, los datos en la integración de usuarios y los modelos de sincronización, necesitarán ser unificados para otras combinaciones de modalidades durante las tareas interactivas realistas, para que puedan establecerse los límites reales para realizar la fusión multimodal.

**Diferencias individuales en la interacción multimodal y sus implicaciones en el diseño**

Cuando los usuarios actúan recíprocamente multimodalmente, pueden existir diferencias individuales grandes en la integración de los modelos. Por ejemplo, el trabajo empírico en integración de pluma/voz multimodal, ha revelado dos tipos principales de usuario, los primeros que habitualmente entregan su habla y signos de la pluma solapados o de manera simultánea, y otros que sincronizan signos secuencialmente con pluma y habla. El modelo de la integración dominante de estos usuarios, podría haber sido identificado cuando ellos empezaron a actuar recíprocamente con el sistema.

Es decir, el modelo de la integración de cada usuario, fue establecido desde un inicio y permanecía consistente, aunque se observaron dos modelos de integración distintos entre diferentes usuarios. También, se han observado diferencias sustanciales en la forma de que las personas proceden sobre el lenguaje, ya que son de diferentes grupos lingüísticos, como chinos, españoles, ingleses, etc.

Todos estos puntos encontrados, implican que el sistema multimodal del futuro, sea capaz de adaptar las limitaciones temporales para grupos de usuario diferentes. Con esto, podrían lograr reconocimiento mayor, exactitud y velocidad interactiva. También se han documentado diferencias individuales y culturales entre los usuarios en los modelos de integración. Por ejemplo, las diferencias individuales sustanciales que se tienen en la sincronización temporal entre el hablar y los movimientos del labio. Además, los movimientos de los labios mientras la persona habla son menos exagerados entre los japoneses que los españoles.

Estos resultados, tienen implicaciones para el diseño y el valor esperado del rendimiento de multimedios audio-visual para usuarios diferentes, los cuales se agrupan en interfaces diferentes.

Se debe tener un modelo, que sincronice de manera general las combinaciones de modalidades para cualquier tipo de usuario, no importando cultura o comportamientos individuales que puedan segmentar el desempeño del sistema. De igual forma, esta segmentación debe ser transparente para el usuario y debe proveer optimización para el sistema.

**Características principales del lenguaje multimodal**

Los canales de comunicación, pueden ser tremendamente influyentes transformando el idioma transmitido dentro de ellos. Ahora, las características lingüísticas del idioma multimodal, son cualitativamente muy diferentes de lo hablado o de lo formalmente textual del idioma. De hecho, puede diferir en características tan básicas como la brevedad, volumen semántico, sintáctico, la complejidad, orden de las palabras, el grado de ambigüedad refiriéndose a expresiones, especificación y lingüísticas indirectas.

En muchos aspectos, la multimodalidad del idioma es lingüísticamente más simple que el mismo idioma hablado. En particular, los usuarios completan una misma tarea comunicando menos palabras, siendo más breve, utilizando frases cortas, y menos descripciones complejas cuando está involucrada la multimodalidad, comparado con usar un solo discurso.

Esta situación, ocurre principalmente porque las personas tienen dificultad al hablar sobre el espacio. En un interfaz multimodal flexible, ellos usan una entrada por medio de una pluma electrónica para expresar la información

espacial a fin de evitar usar la voz. La investigación y esfuerzos futuros, también se necesitan en los tipos diferentes de comunicación multimodal, y en otros dominios de la aplicación, para que la generalidad de diferencias de idioma multimodal previamente identificadas, puedan contemplarse y explorarse. Para concluir este tema, en el cual se discute la ciencia cognoscitiva en el diseño de interfaces multimodales, se destacan los puntos siguientes derivados de la discusión anterior. Estos puntos comunes sobre la interacción multimodal incluyen lo siguiente:

- Al construir un sistema multimodal, los usuarios actuarán recíprocamente multimodalmente.
- Hablar y apuntar es el modelo de integración multimodal dominante.
- Las entradas multimodales involucran signos simultáneos.
- El hablar es la modalidad de entrada primaria en cualquier sistema multimodal que la incluya.
- El idioma multimodal no difiere lingüísticamente del idioma unimodal. La integración multimodal involucra redundancia de volumen entre las modalidades.
- Las tecnologías del reconocimiento predispuestas a errores individuales que combinan multimodalidad, pueden producir desconfianza en el funcionamiento.
- Todos los usuarios multimodales, se integran a las ordenes de una manera uniforme.
- Diferentes modalidades de entrada son capaces de transmitir igual volumen de información.
- Reforzar la eficacia es la ventaja principal de los sistemas multimodales.las diversas perspectivas multidisciplinarias
- Una perspectiva multicultural/internacional natural en modelos de comunicación.

En estos 10 puntos, se resume lo expuesto anteriormente sobre la ciencia cognoscitiva y el diseño de las interfaces multimodales, para pasar ahora a la modelación y evaluación de los estilos de interacción en estos sistemas.

### 4.3.2 Modelación y evaluación de estilos de interacción

En la interacción hombre-maquina, la definición de estilos de interacción, es normalmente muy informal. Un estilo de la interacción puede definirse por encima como la manera general que los usuarios actúan recíprocamente con un sistema. Los ejemplos típicos, incluyen manipulación directa, menús y formas de interfaces del usuario gráficas.

La interacción ofrece al usuario opciones diferentes para ejecutar una tarea, o igualmente, ofrece tipos diferentes de acceso a la funcionalidad del sistema. Un estilo de interacción se caracteriza por tres componentes que se describen a continuación:

**Técnicas de la interacción**

Una técnica de la interacción, se define como la manera que se usan los dispositivos de entrada físicos para facilitar acciones del usuario. De igual manera, se utilizan dispositivos de rendimiento para presentar reacciones del sistema. Ofrece mecanismos de mando físicos ofrecidos al usuario y las técnicas de la presentación usadas por el sistema.

**Estructura de la interacción**

La estructura de la interacción se define como la sucesión de acciones del usuario y reacciones del sistema. Refleja la organización de opciones cuando

éstas se presentan al usuario. La estructura de esta sucesión puede ser clasificada como un procedimiento operacional.

**Funcionamientos conceptuales**

Un funcionamiento conceptual, se define como una acción primitiva que el usuario o el sistema puede ejecutar dentro de una interacción dada. Las tareas del usuario, describen lo que el usuario puede hacer con el sistema sin realmente prescribirlo. Basado en el flujo de información, los tipos principales siguientes de funcionamientos conceptuales pueden ser distinguidos como manipulación, percepción y comunicación.

A nivel de evaluación de estilos de interacción, se utiliza efectividad y eficacia en la tarea tomando indicadores de utilidad cuantitativas, mientras se usan valuaciones de usuario como indicadores de utilidad cualitativas. Los indicadores cuantitativos son operacionales en el diseño experimental mientras los cualitativos se recogen en encuestas post-experimentales.

Dos tipos de estilos de interacción se pueden evaluar. Un estilo con un dispositivo que funciona con un telemando y un estilo de la manipulación directa. Se recogen los funcionamientos conceptuales iguales para contrastar ambos tipos de estilos de interacción mientras la estructura y la técnica es variada por ambos estilos de la interacción.

Los resultados estadísticos de un potencial experimento como el planteado anteriormente, hacen pensar en la presencia de un efecto de aprendizaje para ambos estilos de interacción. Desde que la técnica y estructura de la interacción es distinta para los dos tipos de estilos de la interacción, el aprendizaje sugiere un efecto que puede atribuirse a los

funcionamientos conceptuales fijos para ambos estilos de la interacción. En el contexto de este diseño, se sugiere que los usuarios construyan un modelo conceptual, para que el trabajo con el estilo de telemando pueda usarse de igual forma el de manipulación directa.

Con tal modelo, los usuarios pueden hacer menos errores al trabajar con el estilo de manipulación directa, que usuarios que no usaron el estilo de telemando. Otra observación interesante en las evaluaciones, es el hecho de verificar que los usuarios no juzguen por la presentación ya que esto no es significante. Las diferencias significantes entre ambos estilos de interacción pueden atribuirse a las características cuantificadas de su aplicación.

Al mismo tiempo de la evaluación de los estilos de interacción, se debe tener el método adecuado para la implementación del sistema, lo cual se plantea a continuación.

### 4.3.3 Métodos e información para diseño

El diseño de nuevos sistemas multimodales, ha estado inspirado y organizado grandemente por dos cosas. Primero, la ciencia cognoscitiva en la percepción sensorial y coordinación de modalidades durante la producción de un diseño de usuario, también como la información sobre qué sistemas debe reconocer y cómo las arquitecturas multimodales deben estar organizadas.

Dado lo complejo de la naturaleza de los usuarios en la interacción multimodal, la ciencia cognoscitiva tiene y continuará jugando un papel esencial guiando los diseños de sistemas multimodales robustos. En este aspecto, la perspectiva de la multidisciplina es fundamental en el diseño de sistemas multimodales exitosos con una interfaz gráfica de usuario asociada.

Las simulaciones automáticas de alta fidelidad, también han jugado un papel importante en la creación de prototipos nuevos de sistemas multimodales. Cuando un nuevo sistema multimodal está en las fases de la planificación y diseño del sistema, se utiliza estas simulaciones para visualizar el nuevo sistema y planear el flujo inicial secuencial de la interacción hombre-máquina. Estos diseños provisionales, planean una simulación de sistema multimodal, dando una colección de datos considerable.

Durante la simulación se prueba la alta fidelidad; un usuario actúa recíprocamente con lo que se cree es un sistema multimodal totalmente funcional, aunque la interfaz es realmente una presentación simulada y que no permitiría que el sistema respondiera en forma totalmente funcional. Durante la interacción, el sistema recibe la información de un ayudante de programador.

Cuando el usuario actúa recíprocamente inicialmente, el programador rastrea su entrada multimodal y proporciona respuestas del sistema rápidamente y con precisión como posible. Apoyado en esta técnica, el programador hace uso de una simulación automatizada como si él fuera el software que se diseñará, esto, para apoyar la velocidad interactiva, el realismo con respecto al objetivo del sistema y otras características importantes.

Por ejemplo, con estas herramientas automatizadas, el programador puede hacer una sola selección en un campo de la estación de trabajo enviar las respuestas rápidamente simulando el sistema al usuario durante una sesión con cierta colección de datos. Las simulaciones de alta fidelidad han sido el método preferido para los prototipos de sistemas multimodales por varias razones.

Las simulaciones son relativamente fáciles y baratas de adaptar, comparado con construir un sistema completo. También permiten alterar un sistema planeado con características mayores a las iniciales y para estudiar el impacto de las diferentes características de la interfaz de una manera sistemática y científica.

En comparación, un sistema particular con sus características fijas es menos flexible. Usando técnicas de simulación, los diseños se adaptan de manera rápida y permite a los desarrolladores ver con un sentido más amplio las características inicialmente propuestas.

En un sentido práctico, la investigación de la simulación puede ayudar en la evaluación de los diseños, ya que se pueden tomar decisiones sobre los mismos diseños del sistema y buscar alternativas para fortalecer estos sistemas y hacerlos mas utilizables.

Para el apoyo del desarrollo y comercialización de sistemas multimodales, adicionalmente se necesita infraestructura que incluye: las herramientas de simulación para construir rápidamente los diseños o reconfigurarlos, las herramientas automatizadas para coleccionar y analizar entornos multimodales y las herramientas automatizadas para interacción de los nuevos sistemas multimodales para mejorar su rendimiento.

## 4.4 Integración de tarea de reconocimiento de voz

El primer sistema de reconocimiento de voz, fue desarrollado en 1952 sobre una computadora analógica, usando voz discretizada para reconocer los dígitos del 0 al 9 con un algoritmo de plantilla de concordancia dependiente de la persona que habla. Mas tarde en esa misma década, un sistema con

atributos similares fue desarrollado que reconoció consonantes y vocales. En los años sesenta la investigación en reconocimiento de voz se movió a las computadoras digitales. Esta plataforma proporcionó las bases para la tecnología de reconocimiento de voz como se conoce hoy en día.

A pesar del rápido progreso inicial, las limitaciones en arquitecturas de computadoras, previno cualquier desarrollo comercial de sistemas de reconocimiento de voz. En la última década, sin embargo, un número de sistemas comerciales han sido exitosamente desarrollados. A pesar de estos avances, el verdadero procesamiento de voz espera aun varios años por venir. Por lo tanto, un sistema exitoso conducido por voz, debe tener en cuenta las limitaciones de la tecnología actual. Estas limitaciones incluyen la dependencia de la persona que habla, la continuidad de voz y el tamaño del vocabulario.

Los sistemas independientes de la persona que habla, pueden reconocer voz de cualquier persona. Los sistemas dependientes de la persona que habla deben ser entrenados para cada usuario individual, pero típicamente tienen más altas tasas de exactitud. Los sistemas adaptables a la persona que habla, un enfoque híbrido, inicia con plantillas independientes de la persona que habla y las adapta a usuarios específicos sobre el tiempo sin entrenamiento explícito.

Los sistemas de voz continuos pueden reconocer palabras habladas en un ritmo natural mientras que los sistemas de palabras aisladas requieren de una pausa deliberada entre cada palabra. No obstante más deseable, la voz continua es más difícil de procesar por la dificultad en detectar los límites de cada palabra.

Los grandes vocabularios causan dificultades en mantener exactitud, pero los pequeños pueden imponer restricciones no deseadas sobre la naturalidad

de la comunicación. A menudo el vocabulario debe ser restringido por reglas gramaticales las cuales identifican como las palabras pueden ser habladas en el contexto.

Junto con las características técnicas de sistemas de reconocimiento de voz, es importante entender los factores humanos de voz como una modalidad de la interfaz. La más significante es que la voz es temporal. Una vez pronunciada la información, ya no se dispone más de ella. Esto puede representar una carga adicional para la memoria del usuario y limita severamente la habilidad de repasar, revisar y verificar la información de referencias cruzadas. La voz puede ser usada a distancia lo cual la hace ideal para situaciones de manos y ojos ocupados. Es omnidireccional y por lo tanto puede comunicarse a múltiples usuarios. Sin embargo, esto tiene implicaciones relacionadas a la privacidad y la seguridad.

Por otra parte, más que otras modalidades, hay la posibilidad de antropomorfismo cuando se usa el reconocimiento de voz. Ha sido documentado que los usuarios tienden a sobrestimar las capacidades de un sistema si una interfaz de voz es usada y que los usuarios son más tentados a tratar el dispositivo como otra persona.

El uso del ratón para eliminar la ambigüedad de la entrada del contexto de voz también ha sido explorado. Ejemplificando ese tema, la compañía Boeing para el Sistema de Control y Alerta de Vuelos *(Airborne Warning and Control System*, AWACS)[19]. Notando que la comunicación humana es multidimensional y que las conversaciones incluyen más que palabras habladas, se utilizó una combinación de datos gráficos y verbales donde uno completa o elimina la

[19] Referencia bibliográfica [9] Airborne warning and control system.

ambigüedad del otro. Dentro de este marco de referencia, los operadores podrían hacer requerimientos hablando comandos mientras que simultáneamente, seleccionan objetos gráficos con un ratón para determinar el contexto de estos comandos.

Un enfoque similar fue tomado mientras se integró una interfaz de lenguaje natural con una herramienta de planeación de prueba de alerta de vuelos gráfica en sus etapas iniciales en el Laboratorio de Investigación Naval de los Estados Unidos. El uso de lenguaje natural proporcionó poder expresivo sobre y más allá de lo que es posible con la manipulación directa. Por ejemplo, usando voz, un usuario pudo especificar un comando, una referencia y un destino, tal como "Mover el atacante 14 a la estación 5". Alternativamente, usando una entrada multimodal, un usuario pudo especificar el comando y la referencia solamente como "Mover el atacante 14". El destino pudo haber sido seleccionado entonces usando el ratón al apuntar sobre una posición de un mapa gráfico.

Siguiendo guías intuitivas, estos esfuerzos parecieron integrar voz en tareas de entradas multimodales y multidimensionales cuando los atributos de entrada fueron perceptualmente separables. Ejemplos de lo anterior, es cuando hubo un cambio en el contexto o función, tal como una identificación de referencia contra datos de entrada, descripción contra examen en el contexto de tiempo, y datos de entrada contra comandos. Esto sugiere que en una evaluación empírica, el desempeño puede mejorar cuando los atributos perceptualmente separables son entradas usando diferentes modalidades.

Para finalizar, se ha presentado una perspectiva sobre la multimodalidad y la voz. El hablar, es la forma más común de expresión que existe y en la integración de tareas multimodales, es la modalidad de entrada primaria en

cualquier sistema multimodal que la incluya. Se ha definido el diseño, modelación y evaluación de éstas interfaces, para finalizar en la integración de la tarea, específicamente la voz. Esto da la pauta, para plantear, diseñar, evaluar e integrar la voz en una interfaz hombre-máquina multimodal, a fin de que esta interfaz sea eficiente.

# 5. EXPERIMENTO DE TAREA EN UN PROCESO UNIMODAL FRENTE A UN PROCESO MULTIMODAL

Los sistemas multimodales representan un paso hacia la comunicación natural con los sistemas computacionales. Por ejemplo, la computadora es capaz de reconocer a una persona que habla, entender lo que dice y generar la salida apropiada para el usuario.

Dentro de las modalidades que actualmente se utilizan en los sistemas multimodales se encuentra la voz. Esta modalidad, aunque no se ha perfeccionado del todo, tiene aspectos importantes que son aprovechados en la multimodalidad. El objetivo principal de los sistemas de reconocimiento de voz, es desarrollar interfaces centradas en las necesidades del usuario, aprovechando una de las capacidades que tiene el hombre para comunicarse, la expresión oral.

Estos sistemas han probado su utilidad para ciertas aplicaciones. Uno de los medios más populares para las aplicaciones de voz es el teléfono por razones de facilidad de uso, disponibilidad y costo accesible.

Las aplicaciones basadas en este tipo de reconocedores son servicios financieros, asistencia de directorio, llamadas por cobrar (operadora automática), transferencia de llamadas telefónicas, consultas de información (clima, tráfico, reservaciones). Las ventajas que presentan este tipo de aplicaciones son que al interactuar el usuario utiliza la eficiencia del habla (rápida, flexible, natural) y está libre de movimientos de las manos en caso de que las tenga ocupadas. Existen otras aplicaciones que no se basan en el

teléfono, por ejemplo el dictado automático. También el reconocimiento de voz es usado en compañías en donde la entrada de datos o comandos por voz es requerida tales como desarrollo de inventarios, control de robots, etc.

La voz tiene ventajas considerables en la comunicación de los seres humanos, por lo cual, se debe explotar para optimizar tareas que puedan ser ejecutadas por máquinas. Una forma de explotar la voz para optimizar tareas es sobreponer sobre las propias debilidades de la voz, alguna modalidad extra para que las ventajas de esta nueva modalidad compensen las debilidades al hablar.

Con esto, la unión de las tareas se genera un experimento multimodal, el cual tiene que combinar las dos modalidades como entrada en un sistema. El experimento que se presenta en los párrafos siguientes está basado en la filosofía de "apuntar y hablar". Este experimento servirá para determinar las ventajas de trabajar con la multimodalidad versus la unimodalidad.

## 5.1 El experimento

La multimodalidad provee muchas ventajas en las interacciones hombre - máquina y es la que proporciona el valor agregado a estos sistemas en los cuales la interacción con el humano es fundamental. En la constante de intentar modelar en lo posible, el comportamiento humano, se ha planteado la necesidad de buscar las modalidades más eficientes y eficaces con las cuales cuenta el hombre para comunicar sus ideas.

Dentro de las modalidades que llenan las características anteriores y que actualmente ocupa un espacio preponderante en las interacciones hombre-

máquina es la voz. La voz es la forma de comunicación más eficiente con que cuenta el ser humano.

Por la naturaleza del comportamiento de la voz en el hombre, el llegar a perfeccionar esta modalidad para reconocer la voz en el ser humano por medio de una computadora es un tema aun por terminar. De igual manera, el reconocimiento de voz tiene ventajas importantes que deben ser aprovechadas. Para aprovechar estas ventajas se deben debilitar las desventajas del reconocimiento de voz y esto se plantea añadiendo una modalidad extra que realice tal efecto.

Esto permite una interacción con una maquina por medio de la voz, optimizando tareas que actualmente se manejan de una forma directa o unimodal, y que al aplicar esta metodología se verá mejorada la entrada de datos y su interacción con el usuario.

### 5.1.1 Planteamiento del experimento

Con base en los argumentos expuestos anteriormente, el experimento será la unificación de las modalidades de voz con una manipulación directa para una tarea unimodal. La finalidad de este experimento, es la creación de una entrada integral de reconocimiento de voz multimodal, para mejorar la tarea y verificar la optimización de tiempo y funcionamiento de la misma con este método.

Al obtener los resultados de este experimento, se intenta evaluar los beneficios que aporta la integración de tareas en ambientes de reconocimiento de voz multimodal y determinar las características de estas tareas para lograr establecer tareas candidatas para la utilización de este método multimodal.

Para tal efecto, es necesario establecer lo que se pretende con este experimento, es decir, los objetivos.

#### 5.1.1.1 Objetivo general

Determinar el nivel de optimización y aceptación por parte de los usuarios de una tarea creada en una interfaz hombre-maquina con reconocimiento de voz multimodal.

#### 5.1.1.2 Actividades específicas

- Plantear una interfaz de usuario multimodal a partir de una tarea unimodal específica.
- Diseñar una interfaz hombre-máquina utilizando el reconocimiento de voz multimodal.
- Utilizar una herramienta para recolección de datos para establecer la amigabilidad, facilidad, rapidez y adaptabilidad de la interfaz hombre-máquina de reconocimiento de voz multimodal.
- Evaluar los tiempos que se utilizan los diferentes usuarios del experimento multimodal versus el experimento unimodal.

Además del planteamiento inicial y la exposición de los objetivos, es necesario realizar una justificación de este experimento para exponer las razones de la realización del mismo.

### 5.1.2 Justificación

Dado el planteamiento de este trabajo que es la integración de tareas en ambientes de reconocimiento de voz multimodal, se hace necesario el diseñar de manera experimental este método, debido a que ésta técnica, a pesar que ya se tiene cierto tiempo de investigación, no es está depurada y es aún objeto de investigación.

El proponer este experimento, es la verificación y sustentación de los fundamentos teóricos expuestos en los capítulos anteriores. También es comprender de forma gráfica, cuantitativa y cualitativa los beneficios que ésta técnica puede proveer.

El experimento pretende también determinar las características funcionales y las necesidades de los usuarios respecto a interfaces mucho más amigables y cercanas a una relación humana.

Ciertamente, este experimento se realiza bajo condiciones ideales, las cuales se explican posteriormente, por lo cual se presenta como sustento de la teoría. Debido a esto, el trabajo futuro deberá utilizar como cimientos los conceptos expuestos y materializados en la presente experimentación para depurar esta metodología y reforzar esta teoría cada vez más.

## 5.2 Tipo de experimentación

En este experimento se busca desarrollar una imagen o fiel representación (descripción) del caso estudiado a partir de sus características. Describir en este caso es sinónimo de medir. Se medirán variables o conceptos con el fin de especificar las propiedades importantes de la integración de tareas en

ambientes de reconocimiento de voz multimodal que es el tema que se está analizando.

Con base en lo expuesto anteriormente, el tipo de experimentación que se ha seleccionado es el de descripción. Este estudio comprende la descripción, registro, análisis e interpretación de la naturaleza actual, y la composición o procesos de la metodología de reconocimiento de voz multimodal.

La investigación descriptiva trabaja sobre realidades de hecho, y su característica fundamental es la de presentarnos una interpretación correcta. El énfasis está en el estudio independiente de cada característica, es posible que de alguna manera se integren la mediciones de dos o más características con en fin de determinar cómo es o cómo se manifiesta el experimento. Pero en ningún momento se pretende establecer la forma de relación entre estas características.

Básicamente, éste es un estudio descriptivo para analizar como se comporta este tipo de métodos multimodales utilizando el reconocimiento de voz. Este experimento describirá como se manifiestan las ventajas y desventajas de la mutimodalidad y unimodalidad, así como los componentes que en esta metodología se utilizan.

La experimentación descriptiva, trabaja sobre realidades de hecho y su característica fundamental es la de presentar una interpretación correcta. Este experimento incluirá el tipo de estudio de un caso exploratorio de comparación.

Los estudios descriptivos miden de manera más bien independiente los conceptos o variables a los que se refieren. Aunque, desde luego, pueden integrar las mediciones de cada una de dichas variables para decir cómo es y

cómo se manifiesta esta metodología, su objetivo no es indicar cómo se relacionan las variables medidas.

Los resultados de la indagación teórica, indican qué aspectos son importantes o relevantes y cuáles no respecto a éste tema. Pero ante la duda, convendrá seguir manteniendo los factores sospechosos de ser irrelevantes: en todo caso la duda quedará aclarada en esta experimentación descriptiva.

## 5.3 Definición conceptual de las variables

Dados los objetivos planteados previamente y para poder llevar a cabo los mismos, se definen las siguientes variables a ser observadas para este experimento:

**Optimización de la tarea**

**Definición Conceptual**

El concepto de optimización es muy amplio y se puede evaluar desde varios puntos de vista. En principio, se puede definir como la mejor manera de realizar una actividad o tarea. Para ello se debe identificar cómo realmente funcionan las cosas para hacerlas funcionar mejor.

El tiempo, por su parte, es factor fundamental para cualquier actividad, por lo cual, su ahorro viene a ser un factor positivo para cualquier proceso, dando con ello un mejor funcionamiento a ese nivel. Este es el sentido que se le dará a la optimización de la tarea que se observará. El minimizar tiempo para realizar una tarea significará una optimización en el funcionamiento del proceso.

La tarea a observar, en un inicio unimodal y posteriormente transformada a multimodal, se evaluará para verificar la optimización de tiempo una respecto a la otra.

**Definición Operacional**

Las herramientas a utilizar son parte de la estadística descriptiva ya que con ella se puede determinar conclusiones a partir de comparaciones y comportamiento observados para un experimento dado.

Cuando se va a realizar un análisis estadístico de datos es aconsejable realizar previamente una estadística descriptiva de las variables, para obtener información de dichas variables o simplemente para verificar posibles errores en los datos.

La estadística estudia los métodos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones válidas y tomar decisiones razonables basadas en el análisis.

En un sentido menos amplio, el término estadística se usa para denotar los propios datos, o números derivados de ellos, tales como los promedios. Este es un análisis descriptivo de los tiempos de la tarea con interacción unimodal inicialmente, y luego con interacción multimodal.

Con este tipo de análisis se obtiene una estadística descriptiva (media, desviación estándar, máximo, mínimo,…) de las variables

**Indicadores**

- Determinación de tiempos por medio de control interno de la tarea del experimento, esto por medio de la interfaz

- Utilización de estadística descriptiva para comparación de datos del experimento unimodal versus el experimento multimodal.

**Aceptación de interfaz por parte del usuario**

**Definición Conceptual**

Dada la importancia del usuario en cualquier interfaz, se debe evaluar que tan agradable son las interfaces unimodales y multimodales para una misma tarea.

Es fácil caer en la tentación de suponer que una atractiva interfaz gráfica es siempre mejor que cualquier otra interfaz. Es necesario un proceso de evaluación de la calidad de las posibles interfaces, evaluación que ha de tener en cuanta la usabilidad y el cumplimiento de los requerimientos del usuarios.

El sistemático proceso de evaluación del diseño de las interfaces gráficas puede ser costoso y complicado, ya que en muchos casos involucrará a especialistas en ramas muy diversas y el uso de laboratorios preparados a tal efecto. Por ello, este tipo de evaluación no siempre es posible y hay que restringirse a evaluaciones de usabilidad de acuerdo a normas sencillas tendentes a localizar deficiencias de la interfaz.

**Definición Operacional**

La evaluación debe ser específica sobre la reacción y comportamiento del usuario al cambio de interacciones. Se toma la técnica de recolección de datos por medio de la evaluación de puntos importantes en el momento de la experimentación, basándose en la observación y cambios de actitud de los usuarios respecto al cambio de interacción.

De igual forma, se observa la facilidad de utilización y como el usuario ve el mundo real respecto a las interacciones presentadas.

**Indicadores**

- Evaluación, en base a la observación, de puntos importantes respecto a actitud de usuarios
- Comparación por medio de estadística descriptiva de facilidad, rapidez y adaptabilidad del usuario de aprendizaje dados los tiempos obtenidos en el propio experimento.

**Interfaz gráfica hombre-maquina**

**Definición Conceptual**

La interfaz gráfica hombre-máquina es una conexión e interacción entre hardware, software y usuario. Las interfases de hardware son los conectores, cables, etc. que transportan las señales eléctricas en un orden prescrito. Las interfaces de software son los lenguajes, códigos y mensajes que utilizan los programas para comunicarse unos con otros, tal como un programa de aplicación y el sistema operativo.

En este tipo de interfaces se debe diferenciar la parte de interacción de máquina y la del humano. Es complicado unificar estas interacciones ya que hay que modelar la interacción humana con la maquina como si fuese otro humano más.

Es importante determinar, el nivel de “compresión humana” de la máquina, ya que con esto, la interfaz tendrá un mejor desempeño.

**Definición Operacional**

La evaluación debe ser específica sobre el nivel de la comprensión de la interfaz de su ambiente humano. De igual forma, se estudia por medio del ambiente en que se realizará el experimento, la actividad a realizar o la tarea a evaluar con las dos interacciones.

Se debe separar las interacciones, la de usuario y la de máquina, para determinar el comportamiento de los usuarios en estos ambientes.

**Indicadores**

- Determinación de interacciones máquina.
- Determinación de interacciones hombre.
- Desarrollar herramienta y técnica para ayudar a que la tarea que se evaluará sea idónea para las actividades a las cuales se quieran aplicar.-Determinar una interacción eficiente, efectiva y segura.

**Reconocimiento de voz multimodal**

**Definición Conceptual**

El reconocimiento de voz es una técnica aun por perfeccionar, pero se vale de la multimodalidad para cubrir algunos de los puntos negativos con que cuenta. Se necesita tener un nivel de comprensión por medio de la interfaz de la voz, según condiciones de tono de voz, de idioma, de pronunciación, etc.

Al unificar otra tarea al reconocimiento de voz, se debe evaluar el nivel de integración de la misma a esta tecnología, por lo cual, se debe buscar la modalidad apropiada para que el efecto sea positivo.

**Definición Operacional**

En estas circunstancias, se debe evaluar según las condiciones que presente el experimento y los usuarios a ser observados. De esto dependerá la Interfaz de reconocimiento de voz.

Se debe tener presente el método para la unificación de las tareas y la tarea a unificar con el reconocimiento de voz, porque esto definirá la multimodalidad en la tarea unificada

**Indicadores**

- Evaluación del tipo de voz, pronunciación y lenguaje a utilizar.
- Nivel de unificación de tareas multimodales con reconocimiento de voz.

## 5.4 Diseño del experimento

El experimento se realizará para el ingreso de datos a un formulario, en la cual, no se cuente con un orden específico de entrada de los mismos, esto quiere decir, contar con varios campos por ingresar y no tener el orden establecido para llenarlos. Por ejemplo, esto se puede plantear en evaluaciones interactivas en donde el usuario debe estar pendiente de la pantalla y de la información que el sistema le traslade por este medio.

El experimento le deberá indicar donde realizar la acción, la cual inicialmente es unimodal. Posteriormente, el experimento suprime la acción unimodal por una multimodal donde se utiliza la voz.

Esencialmente, el experimento consiste en la respuesta a 12 preguntas generales. La identificación y respuesta a las preguntas debe ser por medio de

la descripción escrita o hablada. Por cada usuario, se toma al azar la versión del programa a utilizar y dependiendo de la versión, así se tomará el funcionamiento del experimento. En general, cada versión tiene un paso específico, los cuales se describen a continuación:

**Experimentación unimodal**

Este módulo es la parte del experimento con manipulación directa, donde se evalúa la interacción que tanga el usuario con el experimento unimodalmente, es decir, apuntando, dando clic y seleccionando, una modalidad después de la otra. En principio, se despliega las instrucciones y posteriormente, se inicia con las preguntas. Cada pregunta es de selección, por lo cual es tienen por cada pregunta la selección de varias opciones. Aquí solo se apunta y se da clic con el ratón. El programa calculará el tiempo de ingreso de datos.

**Experimentación multimodal**

Este modulo es la parte del experimento en ambiente de reconocimiento de voz multimodal, donde se evalúa la interacción que tanga el usuario con el experimento multimodalmente, es decir, apuntando y hablando en una modalidad unificada. Las mismas preguntas que se presentaron en el módulo unimodal, son las preguntas que se presentan en este módulo. El usuario deberá apuntar y hablar como una tarea unificada para responder a las preguntas que se le presentan y dependerá de la pronunciación de las palabras por medio del micrófono el éxito del reconocimiento de la voz y por lógica de las preguntas. De igual forma que en la fase unimodal, el programa calculará el tiempo empleado para la entrada de datos.

### 5.4.1 Materiales a utilizar

- 1 Computador procesador Pentium III, 128 Mb RAM, entrada y salida de audio.
- Sistema Operativo Microsoft® Windows® 98/Me/2000/XP.
- 1 micrófono.
- Altavoces.
- Visual Basic 6.0.
- Microsoft® Access 2000/XP.
- Librería Microsoft® Speech SDK 5.1, SAPI versión 5.1.4324.00 para reconocimento de voz.

### 5.4.2 Escenario del experimento

El experimento debe reunir condiciones ideales para que sea exitoso, o por lo menos, tener una buena probabilidad de resultar exitoso, respecto al ambiente en el que se desenvuelva.

La condiciones ideales a nivel ambiente externo del experimento debe se un lugar donde no exista ningún ruido, donde el usuario este aislado y se concentre en la actividad a realizar.

Los instrumentos a utilizar pueden provocar fallos, por lo cual, el ambiente debe ser como se explico anteriormente para que se reduzca la posibilidad de fallo.

## 5.5 Alcances y límites

### 5.5.1 Alcances

- Aplicar el proceso de ingreso de información a los campos de una base de datos con esta tecnología de interfaz multimodal, utilizando el puntero del ratón para elegir que campo se le enviaran los datos por medio de la voz.
- Realizar experimentos sobre la aplicación y evaluar optimización de tareas específicas del ingreso de información a una base de datos que puedan beneficiarse de una interfaz multimodal en ambiente de reconocimiento de voz.
- Evaluar las técnicas y factibilidades, sobre la tarea en observación a optimizarse por medio de la interfaz multimodal en ambiente de reconocimiento de voz, definiendo las características relevantes de cada una.
- Presentación de evaluación comparativa del aumento de velocidad, exactitud y aceptación de entradas multimodales cuando las tareas sean separadas o integradas.

### 5.5.2 Límites

- Implementación de la aplicación basada en Microsoft® Speech SDK 5.1, por lo cual hay que adaptarse a la implementación con que ya cuenta el software.
- Construcción del experimento según las restricciones que se encuentren a nivel de hardware y del equipo técnico que se utilice para enviar la voz como entrada al sistema.

- Restricción del experimento al ambiente externo, por posibles factores independientes de la aplicación, que pueden influir en errores, por ejemplo, el ruido.
- Inexactitud en herramientas para realizar cálculos y mediciones en la elaboración de las comparaciones estadísticas.
- Condición en el tipo de voz, rapidez, claridad de pronunciación y acento con que se hable, para que el sistema pueda interpretar las instrucciones de lenguaje hablado.
- Orientación de la aplicación únicamente a la verificación de tiempos, exactitud y beneficios que se hayan descubierto durante el experimento.
- Habilidad para la utilización del sistema, tanto en la manipulación directa del puntero del ratón, como en la forma de hablar.

## 5.6 Selección de la muestra

Para la seleccionar las personas que se evaluarán en este experimento, surge un problema: como habitualmente las poblaciones son muy grandes, no se puede obtener los datos de todos los individuos, y entonces se debe seleccionar una parte de esa población.

La muestra es un subgrupo de la población y puede ser probabilístico y no probabilístico. Para este experimento se ha seleccionado una muestra no probabilística, ya que se elegirá a las personas de estudio, debido a la naturaleza de la experimentación.

En el muestreo no probabilístico los individuos de la población no tienen todos la misma probabilidad de aparecer en la muestra, ya que aquí no se hace una estricta selección al azar sino se toman casos típicos que se suponen

representativos de la media poblacional o casos en forma accidental, los que tengan más a mano, etc.

Son aquellos en los que no se puede establecer a priori una probabilidad de los miembros del universo que pueden formar parte de la muestra.

### 5.6.1 Delimitación de la población

La población es constituida por el conjunto de medidas de las variables en estudio, en cada una de las unidades que conforman el universo. Es decir, cada una de las variables en estudio constituye una población que viene dada por el conjunto de valores que ella toma de la realidad que conforman el universo.

Para este experimento, se cuenta con una población N, la cual se define como “individuos que tienen conocimientos computacionales que vive en la ciudad de Guatemala, Guatemala”.

En el caso de este experimento, no es posible tener contacto y observar a todas las unidades de análisis posibles, por lo que es necesario seleccionar un subconjunto de la misma que en efecto represente de manera apropiada a toda la población.

### 5.6.2 Tipo de muestra

El tipo de muestra a utilizar es la no probabilística, ya que se utilizará en forma empírica, es decir, no se efectúa bajo normas probabilística de selección, por lo que en estos procesos que se investigarán intervienen opiniones y criterios personales o no existirá norma bien definida o validada. Se acude a

este tipo de muestra porque es difícil enumerar, listar o precisar el universo objeto de estudio o cuando no existen registros de los datos.

Además, esta selección de los individuos a estudiar será no aleatoria, depende en este caso del experimento. Se utiliza de esta manera, ya que no interesa tanto la población sino ciertas características de las personas en estudio. Se escogerá a los que se ofrecen como voluntarios y que llenen un mínimo de conocimientos respecto al uso de una computadora, pues con esto, estas personas representarán las características que se estudian para este experimento.

Este tipo de muestra no conduce a conclusiones que puedan ser generalizables a la población, este no es el objetivo del experimento, por lo cual se hace la observación pertinente al respecto.

### 5.6.3 Tamaño y obtención de la muestra

Se toma una porción de una población como subconjunto representativo de dicha población. El método muestral no probabilistico, selecciona del marco muestral de la población, una cantidad de individuos no aplicando ninguna formula para calcular el tamaño de muestra.

El objeto del experimento es validar la hipótesis planteada inicialmente y debido a que para la realización de esta actividad se debe contar con circunstancias ideales por la naturaleza del mismo, la muestra será la siguiente: 30 personas organizadas al azar, en 2 grupos de 15 cada uno, independiente del sexo de la persona. Un grupo trabaja el experimento en su versión unimodal y el otro grupo trabaja el experimento en su versión multimodal. Esta discriminación se hace en base a los conocimientos sobre el uso de

computadoras que tenga el individuo, evaluación previa para la discriminación de los grupos.

Se ha hecho la distinción de estos grupos de usuarios debido a las condiciones del experimento. Se evaluará sin distinción de tono de voz y con grupos que el programa irá eligiendo al azar. Esto se realiza con el fin, de no condicionar el experimento y que éste proporcione datos independientes de los usuarios que lo utilicen. Por otra parte, el usuario tiene la incertidumbre de la modalidad a utilizar, por lo cual, se puede evaluar el nivel de adaptación a circunstancias del experimento, ya que se podrá evaluar que tan aceptada es este tipo de interfaces y si tienen buena recepción del usuario.

# 6. RESULTADOS DEL EXPERIMENTO

A partir del experimento planteado en el capitulo anterior, se ha diseñado y realizado la aplicación para efectuar la evaluación que llene los objetivos de esta prueba. El experimento se ha realizado con las especificaciones del planteamiento propuesto y se ha efectuado sobre la muestra de 30 personas, 15 hombres y 15 mujeres. A continuación, se presenta el diseño del experimento.

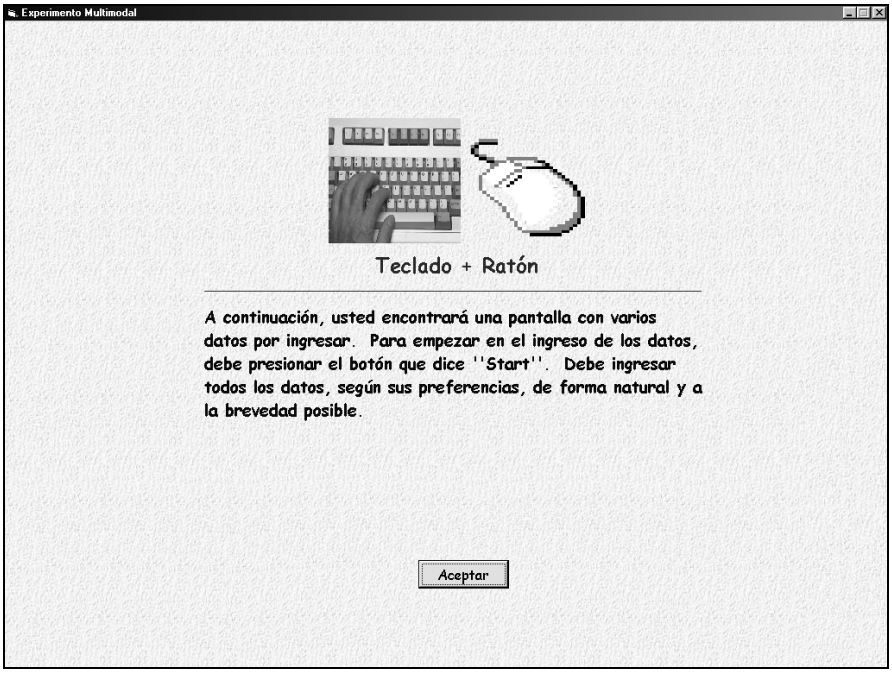
## 6.1 Diseño de la aplicación

La aplicación está formada por 2 estilos de interacción: Unimodal y Multimodal. Los usuarios al utilizar la aplicación, automáticamente se hacen acreedores de su estilo de interacción, acción que realiza la aplicación al azar. La única indicación que se les ha proporcionado a los usuarios para ambos estilos de interacción, es que escojan las opciones instintivamente. No se trata de que tomen cualquier opción, sino que tomen la opción que más les guste a primera vista.

### Estilo de Interacción Unimodal

El estilo Unimodal, es en cual, el usuario debe ingresar los datos al formulario con una manipulación directa, utilizando el puntero del ratón y los clics del mismo. Para este efecto, la aplicación presenta inicialmente una pantalla con las instrucciones de la modalidad que se evaluará. En la figura siguiente se muestra la pantalla de instrucciones.

**Figura 9. Pantalla de instrucciones de experimento unimodal**

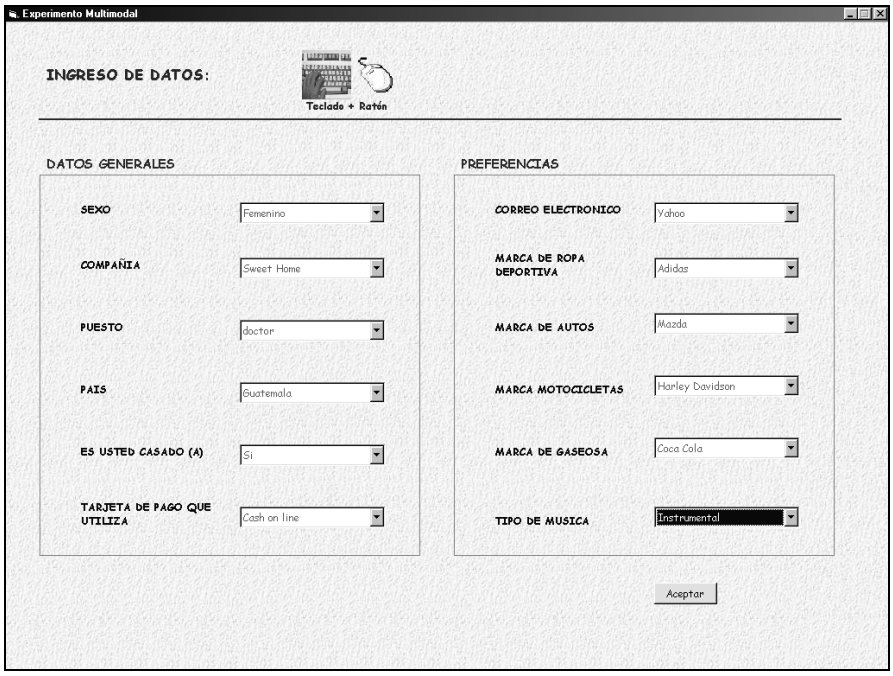



Luego, se presiona el botón **"Aceptar"** y se presentará el estado de inicio para la evaluación unimodal. En este estado de inicio, la pantalla aparece sin datos y se debe presionar el botón **"Empezar"** para inicial la captura de la información.

**Figura 10. Pantalla de instrucciones de experimento unimodal**

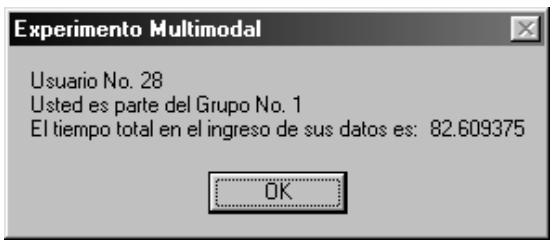

Al presionar el botón, se desplegaran los datos a ingresar y a partir de ese momento, se iniciará la captura de los datos y el conteo del tiempo.

**Figura 11. Pantalla de ingreso de datos experimento unimodal**




Se deben ingresar todos los datos. Cuando se finalice de ingresarlos, se presiona el botón **"Aceptar"** y se verificará si en efecto, están los datos completos. Si están completos los datos, se procede a capturar el tiempo que se utilizó en el ingreso de los mismos y se concluye el experimento para este estilo de interacción. Esto se muestra en una ventana de información, con lo cual, se da por concluido el experimento unimodal. La ventana de información se muestra a continuación.

**Figura 12. Pantalla con los resultados del experimento unimodal**



**Estilo de Interacción Multimodal**

Para el estilo Multimodal, el usuario debe ingresar los datos al formulario con una multimodalidad de “apuntar y hablar”, utilizando el puntero del ratón y su voz, por medio de un micrófono. Para principiar, se despliega una ventana para realizar pruebas de reconocimiento de la voz del usuario, donde aparecen palabras y el usuario debe hablarlas para que el sistema las reconozca.

**Figura 13. Pantalla de entrenamiento para experimento multimodal**


Cuando se reconoce cada palabra, aparece una marca abajo haciendo ver que se ha reconocido la palabra. Este paso se ve gráficamente en la figura 13. Para finalizar el test, se presiona el botón **“Fin Test”.** A continuación, se presenta una ventana con información del experimento multimodal a realizar.

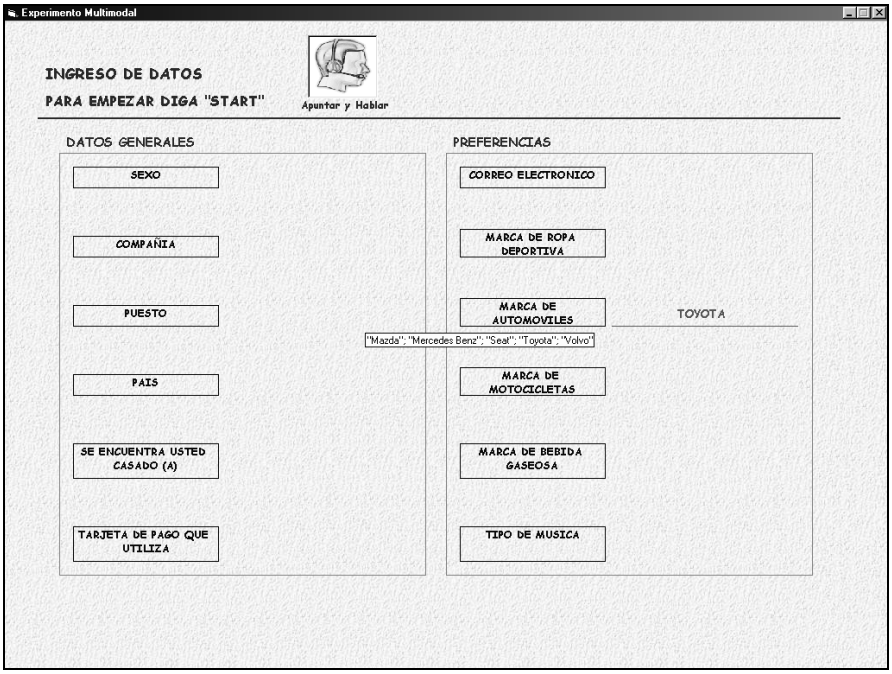
**Figura 14. Pantalla con los resultados del experimento unimodal**


Al terminar de leer las instrucciones, se presiona el botón **"Aceptar"** para continuar. De igual forma que para el estilo Unimodal, la aplicación presenta un estado de inicio para el estilo de interacción Multimodal. En este estado de inicio, la pantalla aparece sin datos y se debe pronunciar la palabra **"Start"** para inicial la captura de los datos, como se muestra en la figura siguiente.

**Figura 15. Pantalla de inicio de experimento multimodal**

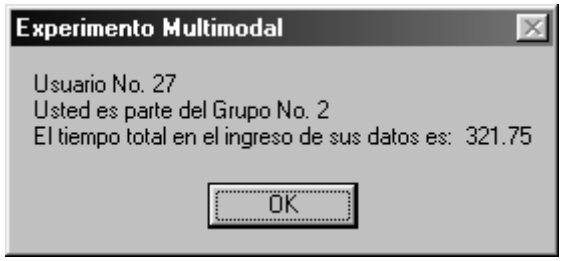
Al pronunciar la palabra **“Start”,** se desplegaran los datos a ingresar y a partir de ese momento, se iniciará la captura de los datos y el conteo del tiempo. El usuario debe apuntar al dato para que muestre las opciones de selección.

**Figura 16. Pantalla de ingreso de datos experimento multimodal**




Se deben ingresar todos los datos. Cuando se finalice de ingresarlos, se presiona pronuncia la palabra **“Finish”** y se verificará si en efecto, están los datos completos. Si están completos los datos, se procede a capturar el tiempo que se utilizó en el ingreso de los mismos y se concluye el experimento para este estilo de interacción. La siguiente figura muestra la ventana de información..

**Figura 17. Pantalla con los resultados del experimento multimodal**




## 6.2 Cálculo estadístico del experimento

En principio, se debe determinar si existen diferencias significativas en los tiempos promedio de ingreso de datos para los dos estilos de interacción. Para esto, se procede a utilizar el modelo estadístico t-Student para poder establecer si existe o no esa diferencia significativa.

### 6.2.1 Modelo t-student

La prueba de t-Student, es un método de análisis estadístico, que compara las medias de dos categorías dentro de una variable dependiente, o las medias de dos grupos diferentes. Es una prueba paramétrica, o sea que solo sirve para comparar variables numéricas de distribución normal.

La prueba t-Student, arroja el valor del estadístico t. Según sea el valor de t, corresponderá un valor de significación estadística determinado. En definitiva la prueba de t-Student contrasta la HP Nula de que la media de la variable numérica “y”, no tiene diferencias para cada grupo de la variable categórica “x”.

En el caso de que se estén estudiando dos variables donde una de ellas es cuantitativa normal considerada como variable respuesta **"Rta"** y la otra variable es dicotómica considerada como variable explicativa **"Exp"**, se pueden aplicar técnicas de estimación por **"IC"** (intervalos de confianza) para diferencia de medias, la prueba t-Student para contrastar la diferencias de medias y técnicas de estimación por IC para el cociente de varianzas.

Para lograr este resultado, la prueba t-Student utiliza los siguientes modelos para el cálculo de sus valores.

**Cálculo de la Estadística Descriptiva**

**Figura 18. Cálculo de los estadísticos descriptivos básicos**

**Cálculo de los estadísticos descriptivos básicos**

Si se denota por $n_1$ y $n_2$ a los tamaños muestrales del primer y del segundo grupos, las medias y las desviaciones típicas para los dos grupos son:

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1}$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2}$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum \left(x_{1i} - \bar{x}_1\right)^2}$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\sum \left(x_{2i} - \bar{x}_2\right)^2}$$

donde $x_{1i}$ indica los valores de la variable Rta para el grupo 1 y $x_{2i}$ indica los valores de la variable Rta para el grupo 2.

**Cálculo de la Estadística Descriptiva**

**Figura 19. Cálculo de intervalo de confianza para la diferencia de medias**

**Cálculo del IC(1 - α)% para la diferencia de medias suponiendo igualdad de varianzas**

Para calcular el IC(1 - α)% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de igualdad de varianzas, tiene la expresión:

$$EE\left(\bar{x}_1 - \bar{x}_2\right) = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

donde $s^2$ recibe el nombre de varianza conjunta ("pooled variance"), que tiene por expresión:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

En segundo lugar para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$, con lo que:

$$IC(1-\alpha)\%\left(\bar{x}_1 - \bar{x}_2\right) = \left[\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{1-\alpha/2,gl}\, EE\left(\bar{x}_1 - \bar{x}_2\right)\right]$$

proporciona el IC buscado.

**Cálculo de t-Student para diferencia de medias**

**Figura 20. Cálculo de t-student para diferencia de medias**

**Cálculo de la prueba t-Student para la diferencia de medias suponiendo igualdad de varianzas**

Para llevar a cabo el contraste:

$H_0$: $\mu_1 - \mu_2 = 0$

$H_1$: $\mu_1 - \mu_2 \neq 0$

suponiendo igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{EE(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$.

### 6.2.2 Prueba t-student para el experimento

El experimento se ha basado en una población N = 30 personas, 15 hombres y 15 mujeres. Cuando se alcanzo la cifra de la población para el experimento, la aplicación posee un módulo para desplegar el resultado

obtenido. En la siguiente figura, se muestra los resultados obtenidos en la ejecución del experimento.

**Figura 21. Resultado global del experimento**


En la figura se muestra un resumen del experimento, donde se segmenta en grupos, por hombres y mujeres, tiempo promedio hombres y mujeres por cada uno de los estilos de interacción, y total de promedio por cada estilo. De igual forma, se presentan 3 graficas con los porcentajes que han ocupada cada uno de los sexos por cada grupo y globalmente. El color rojo, denota el grupo 1, que es la interacción unimodal y el color azul al grupo 2, que es la interacción multimodal.

El detalle de los datos obtenidos para esta prueba, se presentan en la tabla siguiente.

**Tabla I. Resultado total del experimento.**

| Numero | Grupo | Sexo | Tiempo |
|---|---|---|---|
| 1 | 1 | Femenino | 54.796875 |
| 2 | 2 | Masculino | 49.26171875 |
| 3 | 2 | Masculino | 51.41796875 |
| 4 | 1 | Masculino | 58.6875 |
| 5 | 1 | Masculino | 69.79296875 |
| 6 | 2 | Femenino | 58.02734375 |
| 7 | 1 | Femenino | 47.40820313 |
| 8 | 2 | Femenino | 81.015625 |
| 9 | 2 | Femenino | 65.12109375 |
| 10 | 1 | Femenino | 52.1875 |
| 11 | 1 | Femenino | 79.44921875 |
| 12 | 2 | Femenino | 85.46484375 |
| 13 | 1 | Masculino | 53.90625 |
| 14 | 2 | Masculino | 60.5625 |
| 15 | 2 | Masculino | 63.671875 |

| Numero | Grupo | Sexo | Tiempo |
|---|---|---|---|
| 16 | 1 | Masculino | 46.23046875 |
| 17 | 1 | Femenino | 84.08984375 |
| 18 | 1 | Femenino | 60.58984375 |
| 19 | 1 | Femenino | 96.55078125 |
| 20 | 1 | Masculino | 60.75 |
| 21 | 2 | Masculino | 91.48046875 |
| 22 | 2 | Masculino | 73.9921875 |
| 23 | 2 | Femenino | 78.5 |
| 24 | 2 | Masculino | 56.91015625 |
| 25 | 2 | Masculino | 60.53125 |
| 26 | 1 | Femenino | 69.457841 |
| 27 | 1 | Masculino | 64.45124812 |
| 28 | 1 | Femenino | 67.8203125 |
| 29 | 2 | Masculino | 69.1678341 |
| 30 | 2 | Femenino | 71.6857831 |

Sobre esta tabla, se realizan los cálculos para determinar si existe o no una diferencia significativa entre tiempos promedio de ingreso de datos para los dos estilos de interacción.

Inicialmente, se calcula los datos de estadística descriptiva

**Estadística Descriptiva**

**Tabla II. Cálculo estadística descriptiva grupo 1: unimodal**

| **GRUPO 1:** |
|---|
| $\bar{X}1$ **= 64.4113 = Media** |
| **S1 = 14.0026 = Desviación Típica** |

**Tabla III. Cálculo estadística descriptiva grupo 2: multimodal**

| **GRUPO 2:** |
|---|
| $\overline{X2}$ **= 67.7874 = Media** |
| **S2 = 12.4580 = Desviación Típica** |

**Tabla IV. Intervalo de confianza para la diferencia de medias**

| **Intervalo de confianza para la diferencia de medias** |
|---|
| $S^2$ **= 175.63728** |
| $EE(\overline{X1} - \overline{X2})$ **= 4.8392** |

**Tabla V. Student para la diferencia de medias experimiento multimodal**

| **T-Student para diferencia de medias** |
|---|
| $H_0 = U - R <= 0$ **y** $H_1 = U - R > 0$ |
| **t = -0.6977** |
| que bajo la hipótesis nula, sigue una distribución t-Student con grados de libertad gl = (n1-1)+(n2-1) = 28, que tiene asociado un p-valor de 0.4911<br><br>**p-valor 0.4911** |

Calculando el estadístico de la distribución t-Student resulta un valor de 0.4911, con lo cual se sigue que debe aceptarse la hipótesis nula y, por lo tanto, no hay diferencias significativas a favor de algún estilo de interacción determinado.

## 6.3 Evaluación de indicadores

A partir de los objetivos planteados previamente del experimento de un proceso modelo unimodal versus un proceso multimodal y en función del experimento ya realizado, se presenta la evaluación de los indicadores de las variables que se han observado en este proceso de acuerdo a los resultados obtenidos. A continuación, se describe cada uno de estos aspectos.

### 6.3.1 Optimización del proceso

El manejo de la expresión cuantitativa del experimento, es fundamental para determinar si existe o no, una optimización en el proceso realizado. Esta expresión cuantitativa, permite describir el comportamiento del proceso con cada uno de los métodos utilizados para evaluar y comparar. Para definir concretamente este aspecto, se ha definido un mecanismo interno del proceso que toma el tiempo en el que se realiza el experimento, para posteriormente, con ayuda de estadística descriptiva, realizar la comparación de datos. En la sección anterior, se realizó el cálculo estadístico del experimento, obteniendo como resultado, que no existen diferencias significativas a favor de algún estilo de interacción determinado.

En base a este resultado, se puede determinar que el proceso evaluado, no se ha llegado a optimizar, ya que en función de los cálculos realizados, no se determinó alguna diferencia estadística en el promedio de tiempo de ingreso de los datos. La optimización del proceso no ha logrado debido a factores externos que influyen en el rendimiento de la aplicación del reconocimiento de voz multimodal. Estos factores son los siguientes:

**Pronunciación.**

Debido a que la librería utilizada para reconocer la voz, está diseñada para utilizarla en idioma inglés, el nivel de pronunciación y acento de los usuarios fue un elemento que no propició el buen manejo de la aplicación. En ciertos momentos y con ciertos usuarios, la aplicación no realizaba el reconocimiento de las palabras de los usuarios, teniendo que realizar dos o más veces la utilización de la voz para que la aplicación captara el dato.

I**nstrumento de captación de voz.**

El instrumento utilizado para la capturar la voz de los usuarios, entiéndase el micrófono, de igual forma afectó en el nivel de fidelidad de la captación de las palabras pronunciadas por los personas que utilizaron la aplicación. Estos instrumentos, deben poseer un nivel de fidelidad y confianza para que la voz sea captada de forma adecuada.

**Perfiles de usuario para la herramienta de reconocimiento de voz.**

La herramienta para el reconocimiento de voz, la librería SAPI de Microsoft®, debe crear perfiles específicos para cada género (masculino y femenino) y para dos rangos de edades, mayores de 13 años y menores de 13 años. Durante el experimiento, los diferentes tipos y timbres de voz, fueron fundamentales para la selección del perfil del usuario. Se definió un perfil estándar para la evaluación, lo cual provocó que algunos usuarios, tuvieran que entrenar al sistema, según su timbre de voz.

**Ambiente externo.**

El ambiente en donde se realizó el experimento, fue un ambiente donde no existía un silencio absoluto para que la aplicación reconociera la voz con mejor porcentaje de aceptación, debido a que se encontraban ruidos externos, que al tener el micrófono abierto, captaba y distorsionaba la información que interpretaba.

**Inexperiencia de usuario.**

Para algunos usuarios, el cambiar el contexto de comunicación con un computador como lo puede ser hablarle y que éste obedezca, resultó complicado. Ciertos usuarios dejaban ver el que experimentar con la aplicación, sin tomar en cuenta la velocidad con que debían realizar este proceso. Esto provocaba que la naturaleza del experimento, que era manejar el tiempo de ingreso de datos, se afectara al final.

Aparte de estos factores que han afectado el rendimiento de la aplicación, se debe hacer mención de que de igual forma, al no existir diferencias notables, se concluye que el proceso evaluado, estadísticamente, se maneja igual tanto de forma unimodal, como multimodalmente utilizando el reconocimiento de voz, a pesar de todos los factores antes mencionados.

### 6.3.2 Aceptación de interfaz por parte del usuario

Por otra parte, se ha tomado en cuenta la expresión cualitativa del experimento, y esto se conceptualiza en la aceptación de la interfaz por parte del usuario. Con esto, se determina el comportamiento del usuario respecto a la interfaz desarrollada y en general, a la metodología de reconocimiento de voz

multimodal. En función de determinar este nivel de aceptación, se definió un mecanismo de observación de las reacciones de usuario al momento de realizar el experimento. Con esto, se obtiene el grado de valuación que el usuario le da a la aplicación y en general, a la interfaz de reconocimiento de voz multimodal. Estos criterios se han definido en las 3 siguientes reacciones:

**Reacción inicial del usuario.**

Para este experimento, se realizó al azar, la asignación de los usuarios a trabajar con la interacción de reconomiento de voz multimodal y la interacción unimodal. El número total de usuarios que se evaluaron en la metodología de reconocimiento de voz multimodal fueron 15 personas. Derivado de esto, para el 73.33% (11 usuarios) de los usuarios que utilizaron la modalidad de reconocimiento de voz, la primera impresión fue de sorpresa al ver que el computador entendía y obedecía lo que se hablaba. Para el 26.7 % de estos usuarios (4 personas), no les provocó ninguna reacción de sorpresa. Esto se debe a que estos últimos usuarios, tienen más conocimiento de tecnologías como lo puede ser el reconocimiento de voz.

**Reacción durante el experimento.**

Para el grupo de usuarios que no se percibió sorpresa inicial en la interacción de reconocimiento de voz multimodal, se observó que era el grupo que mejor se desenvolvió en la aplicación. Estos usuarios ya cuentan con experiencia en el trabajo de la informática, por lo que les resulto más sencillo el utilizar la aplicación. Para el otro porcentaje de grupo, existieron momentos donde experimentaron dificultades en función de los factores que se expusieron anteriormente en el punto 7.3.1.

**Reacción final y comentarios finales.**

Al finalizar el experimento, el 80 % de los usuarios (12 personas) mostraron su satisfacción de la experiencia vivida, mientras el 20% mostró indiferencia respecto a la aplicación. De igual forma, el 60% (9 usuarios) esbozaron comentarios en función de mejorar la aplicación o nuevas ideas para aplicar.

Al resultar un método nuevo de trabajo, el usuario mostró su aceptación a trabajar de ésta forma, aunque en algunos comentarios, mostraron su escepticismo respecto a que esta metodología se trabaje en un futuro cercano.

# CONCLUSIONES

1. El reconocimiento de voz es una técnica por perfeccionar en busca de modelar el comportamiento humano. El apoyo de la multimodalidad resulta fundamental ya que provee funciones que minimizan las deficiencias de la voz. Estas funciones de igual forma tienen deficiencias, pero al unificar las modalidades, se obtiene una interacción más robusta que permite mejorar procesos unimodales.

2. La observación del experimento, muestra aceptación de los usuarios a una interfaz gráfica multimodal de reconocimiento de voz, pero de igual manera, el utilizar desde siempre la unimodalidad, como lo presentado en una interacción de un ratón, ha generado una resistencia al cambio de interfaz.

3. La voz, por ser un medio de los más eficientes con que el humano puede comunicarse, será un factor importante de llegar a modelarse con buen porcentaje de efectividad. El estudio de la voz como entrada multimodal, proporciona un cambio al paradigma del reconocimiento de voz puro y permite explorar otras áreas fuera propiamente de una interfaz con una computadora.

4. La importancia de la Interfaz gráfica de usuario, se incrementa para este tipo de metodología, ya que depende en buen porcentaje de la eficiencia de la misma el atraer la atención del usuario, y que éste se sienta cómodo con la interacción. Los factores de relevancia en la interfaz son

la amigabilidad, facilidad, rapidez y adaptación del usuario al ambiente de reconocimiento de voz multimodal.

5. La multimodalidad es un concepto muy rico y amplio en su campo de estudio, no se circunscribe a una interacción hombre-máquina, sino que se utiliza en varios campos como la psicología, la medicina, métodos educativos, etc, por lo cual, el ingreso de la multimodalidad, incluyendo el reconocimiento de voz, se ve incrementado no sólo a nivel de procesos informáticos.

# RECOMENDACIONES

1. Debe hacerse mucho énfasis en investigaciones más profundas para poner en práctica esta metodología multimodal, conociendo cuándo la voz es mejor y dónde, y cuándo la información visual sobre el despliegue es mejor. Las dos tienen características muy importantes, y deben ser presentadas de una manera clara y atractiva, ya que la mayoria de usuarios no están familiarizados a tecnología de la voz.

2. A nivel del experimento expuesto, únicamente se presenta como material de prueba de demostración. Es conveniente realizar un estudio a nivel de usuario apropiado para cuando el sistema se desarrolle como un proyecto a largo plazo. Esto hace que la tarea que se desee optimizar con esta metodología, entre en el análisis en conjunto con el usuario. Dependiendo de la naturaleza de la tarea, se puede modelar el sistema con el fin de mejorarlo. Parte del análisis a realizar, es la interfaz de usuario a implementar, debido a que por contar con más opciones, el sistema no necesariamente es mejor.

3. La aplicación del concepto de reconocimiento de voz multimodal, se ha hecho necesaria en varias áreas. Actualmente, ya se está desarrollando sistemas multimodales para web, sistemas multimodales de reconocimiento de voz para no videntes, sistemas de consulta por medio de voz utilizando el teléfono, etc. Dadas las características de este tipo de tareas, es importante explotar los beneficios de esta metodología.

4. Profundizar a nivel de la carrera de ingeniería de ciencias y sistemas este tema, en los cursos de inteligencia artificial y sistemas operativos, a fin de crear conciencia en los estudiantes y despertar la iniciativa de investigación en torno a este contexto. Incluso, el tipo de experimentación en relación a este trabajo de graduación es importante en la carrera, ya que es parte de la actualidad tecnológica, que aunque en este medio no se tengan instrumentos ideales, si es importante para los estudiantes el cimentar conceptos a este nivel para su futuro profesional.

5. Algunos de los procesos administrativos actuales en la Universidad de San Carlos de Guatemala, pueden ser candidatos para esta metodología de reconocimiento de voz multimodal. Por ejemplo, la consulta de libros en la biblioteca utilizando el tacto y la voz, el ingreso de personal a una oficina utilizando la voz y un teclado de números, control de asignaciones de cursos en la universidad por medio de un lápiz electrónico y la voz, etc. Por supuesto, se debe trabajar en perfeccionar el reconocimiento de voz para desarrollar este tipo de metodología, pero queda como recomendación para futuros trabajos el proponer tareas candidatas e implementarlas.

6. Aplicar el concepto de multimodalidad en otras áreas de estudio, no necesariamente en la informática, como por ejemplo la medicina, la psicología, métodos educativos, etc. y estudiar la factibilidad de implementar sistemas donde el reconocimiento de voz se aplique. Es importante estos estudios de factibilidad, ya que no cualquier proceso aplica para esta metodología, y de alguna forma, el aplicar esta técnica puede verse influido negativamente al tratar de implementarlo.

# BIBLIOGRAFÍA


1. Fernández de la Torriente, Gastón. **Cómo hablar correctamente en público.** Editorial Playor. Madrid, España, 1985.

2. Hervás Fernández, Gloria. **La comunicación verbal y no verbal.** Editorial Playor. Madrid, España, 1998.

3. Morris Mano, M. **Arquitectura de Computadores y Ensambladores**. Editorial Prentice Hall. México, 1996.

4. Morris Mano, M. **Sistemas Digitales.** Editorial Prentice Hall, México, 1996.

5. Sotillo, Maria. **Sistemas alternativos de comunicación.** Editorial Trotta, Madrid, España, 1993.

6. Stuart Rusell, Peter Norvig. **Inteligencia Artificial: Un enfoque moderno.** Editorial Prentice Hall. México, 1996.

7. Winston, Patrick Henry Winston. **Inteligencia Artificial.** Editorial Addison-Wesley Iberoamérica, 1994.

BIBLIOGRAFÍA ELECTRÓNICA


8. Agudelo Guzman Marcos, Toro Restrepo Alexander, Casas Herrera Paula Andrea. **Redes Neuronales Artificiales.** http://members.es.tripod.de/recdevoz/pagina_n25.htm. Noviembre 2005

9. **Airborne Warning and Control System**. http://www.boeing.com/defense-space/infoelect/awacs/. Febrero 2006.

10. Ara V. Nefian, Lu Hong Liang, Xiao Xing Liu, Xiaobo Pi. **Visual Interactivity: Audio-Visual Speech Recognition.** http://www.intel.com/technology/computing/applications/avcsr.htm. Febrero 2006.

11. Armengol Torres, Sabaté. **Interfaces de Usuario** http://personals.ip.ictonline.es/+atorres/docs/gui/gui.htm. Julio 2005.

12. Arne Jonsson, Nils Dahlback, Annika Flycht-Ericksson, Pernilla Qvarfordt **Multimodal Dialogue Systems For Industrial Applications.** http://www.ida.liu.se/~arnjo/ceniit.html. Julio 2005.

13. **BioId.** www.bioid.com. Febrero 2006.

14. Ceballos Lizarraga Lizyen, Martínez Castañeda Elizabeth, Young Suárez Janette L. **Sistema orientado a reconocimiento de voz para múltiples aplcaciones (Parte I).** http://proton.ucting.udg.mx/expodec/abr99/cc01/cc01.html. Junio 2005.

15. **CSLU Toolkit**. http://cslu.cse.ogi.edu/toolkit/. Febrero 2006.

16. **Dragon Naturally Speaking**. www.dragontalk.com/NATURAL.htm. Febrero 2006.

17. **Fusión de imagen multimodalidad.** http://hggm.es/imagen/fusion_imagen_multimodal.php. Febrero 2006.

18. **GloveGRASP Gesture Recognition.** http://www.hitl.washington.edu/research/multimodal/GestureGRASP.html. Febrero 2006

19. Grasso Michael A., Finin Tim. **Ambientes de Reconocimiento de Voz Multimodales.** http://www.acm.org/crossroads/espanol/xrds3-3/taskint.html. Febrero 2006.

20. **Hand Motion Gesture Recognition System** www.hitl.washington.edu/research/multimodal/HMRS.html. Febrero 2006.

21. **Intelligent Conversational Avatar.** www.hitl.washington.edu/research/multimodal/avatar.html. Junio 2005

22. Iriarte, Carlos Mauricio. **Reconocimiento de Voz.** http://www.multimania.com/cmib/reconocimientovoz.shtml. Junio 2005.

23. Lazarus, Arnold A. **Evaluación y terapia multimodal.** www.psicologia-online.com/ESMUbeda/multimodal.htm. Febrero 2006.

24. Leal Bustamente, Fernando**. Interface Multimodal. Bibliografía y referencias relativos a Interfaces Multimodales.** http://campus.mor.itesm.mx/~interfaz/articulos.html**.** Julio 2005.

25. Noguera Oliver, López-Polín Herranz Cristina y Salinas Ibáñez Jesús. **El interfaz de usuario.** http://www.filos.unam.mx/POSGRADO/seminarios/pag_robertp/paginas/interfaz.htm. Febrero 2006.

26. **Open source computer vision.** http://www.intel.com/technology/computing/opencv/index.htm. Febrero 2006.

27. Pértega Díaz S., Pita Fernández S. **Métodos paramétricos para la comparación de dos medias. t de Student.** http://www.fisterra.com/mbe/investiga/t_student/t_student.htm. Febrero 2006.

28. **Puerto de Barcelona, España.** www.apb.es. Febrero 2006.

29. **Quickset**. http://cse.ogi.edu/CHCC/QuickSet/mainProj.html. Febrero 2006.

30. Salas Rojas Roberto, Padilla Arciga. **Sistema orientado a reconocimiento de voz para múltiples aplcaciones (Parte II).** http://proton.ucting.udg.mx/expodec/abr99/cc02/cc02.html. Julio 2005.

31. Sharon Oviatt, Antonella DeAngeli & Karen Kuhn. Integration and **Synchronization of Input Modes during Multimodal Human-ComputerInteraction.** http://www.cse.ogi.edu/CHCC/Papers/sharonPaper/Slo/Slo.html. Junio 2005.

32. **Speech SDK 5.1 for Windows® applications.** http://www.microsoft.com/speech/download/sdk51/. Noviembre 2005.

33. **Spoken Language Understanding.** http://cslu.cse.ogi.edu/HLTsurvey/ch1node10.html. Febrero 2006.

34. **Universidad Carnegie Mellon, sección de proyectos, Proyectos Multimodales**. www.is.cs.cmu.edu/mie/. Febrero 2006.

35. Universidad de las Américas – Puebla. **Herramientas del CSLU Toolkit.** http://webserver.pue.udlap.mx/~tesis/ahuactzin_l_a/capitulo_2.html. Junio 2005.

36. **Via Voice.** http://www-306.ibm.com/software/voice/viavoice/. Febrero 2006.

# APÉNDICE

## CÓDIGO FUENTE DEL EXPERIMENTO DE TAREA EN UN PROCESO FRENTE A UN PROCESO MULTIMODAL

**Forma inicial del experimento, identificación y selección al azar del grupo.**

```
Private WithEvents rst As ADODB.Recordset
Private WithEvents rst1 As ADODB.Recordset
Dim MiGrupo As Integer
Private Sub Accesar(MiGrupo As Integer)
   UsuarioActual.TotalForma = 0
   UsuarioActual.Grupo = MiGrupo
   Unload Me
   If MiGrupo = 1 Then Instrucciones.Show Else Test.Show
End Sub
Function ContarRegistros(rst As ADODB.Recordset) As Integer
   ContarRegistros = 0
   Do While Not (rst.EOF)
      ContarRegistros = ContarRegistros + 1
      rst.MoveNext
   Loop
End Function
Private Sub Command1_Click()
   Unload Me
   Call Accesar(MiGrupo)
```

```
End Sub
Private Sub Command2_Click()
   Unload Me
End Sub
Private Sub Command3_Click()
   Unload Me
   Resultados.Show
End Sub
Private Sub Form_Load()
   WebBrowser1.Navigate App.Path & "\imagenes\logo.gif"
   Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
   Randomize
   MiGrupo = Int((2 * Rnd) + 1)
   sBase = App.Path & "\Datos\Datos.mdb"
   Set cnn = New ADODB.Connection
   Set rst = New ADODB.Recordset
   cnn.Open "Provider=Microsoft.Jet.OLEDB.4.0; Data Source=" & sBase
   rst.Open "SELECT * FROM Usuario where grupo = " & Str(MiGrupo), cnn,
adOpenDynamic, adLockOptimistic
   If ContarRegistros(rst) = 15 Then
      If MiGrupo = 1 Then MiGrupo = 2 Else MiGrupo = 1
      rst.Close
      rst.Open "SELECT * FROM Usuario where grupo = " & Str(MiGrupo), cnn,
adOpenDynamic, adLockOptimistic
      If ContarRegistros(rst) = 15 Then
         Command1.Enabled = False
      End If
   End If
   rst.Close
```

```
    Set rst1 = New ADODB.Recordset
    rst.Open "SELECT * FROM Usuario where grupo = 1", cnn, adOpenDynamic,
adLockOptimistic
    rst1.Open "SELECT * FROM Usuario where grupo = 2", cnn,
adOpenDynamic, adLockOptimistic
    If Not (ContarRegistros(rst) > 0 And ContarRegistros(rst1) > 0) Then
        Command3.Enabled = False
    End If
    rst.Close
    rst1.Close
    cnn.Close
    Set rst = Nothing
    Set rst1 = Nothing
    Set cnn = Nothing
End Sub
```

**Desplegar instrucciones por cada uno de los grupos de trabajo.**

```
Private Sub Command1_Click()
If UsuarioActual.Grupo = 1 Then
    Unload Me
    Grupo1.Show
Else
    Unload Me
    Grupo2.Show
End If
End Sub
Private Sub Form_Load()
    Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
```

```
    If UsuarioActual.Grupo = 1 Then DesplegarGrupo1 Else DesplegarGrupo2
End Sub
Private Sub DesplegarGrupo1()
    Image1.Picture = LoadPicture(App.Path & "\imagenes\mouse.gif")
    Image2.Picture = LoadPicture(App.Path & "\imagenes\teclado.jpg")
    Image3.Visible = False
    Label1.Caption = "Teclado + Ratón"
    Label2.Caption = "A continuación, usted encontrará una pantalla con varios " & _  "datos por ingresar.  Para empezar en el ingreso de los datos, " & _ "debe presionar el botón que dice "Start".  Debe ingresar todos " & _ "los datos, según sus preferencias, de forma natural y " & _  "a la brevedad posible.  "
End Sub
Private Sub DesplegarGrupo2()
    Image3.Picture = LoadPicture(App.Path & "\imagenes\voz.bmp")
    Image2.Visible = False
    Image1.Visible = False
    Label1.Caption = "Apuntar y Hablar"
    Label2.Caption = "A continuación, usted encontrará una pantalla con varios " & _  "datos por ingresar.  Para empezar el ingreso de los datos, " & _ "debe pronunciar la palabra "Start".  Para saber las opciones " & _ "por cada uno de los datos, se posiciona en el nombre del " & _ "dato y se desplegarán las opciones.  Para ingresar los " & _ "datos, usted debe posicionarse con el puntero del ratón " & _  "en el dato que se le requiere y pronunciar la opción que " & _  "prefiera. Debe ingresar los datos de forma natural y a la brevedad " & _ "posible.  Cuando finalice el ingreso, pronuncie " & _  "la palabra "Finish"."
End Sub
```

**Forma de trabajo de participantes del grupo No. 1.**

```
Dim StartT, FinishT, TotalTimeT
Private WithEvents rst As ADODB.Recordset
Private Sub Command1_Click()
Dim numero As Integer
If (Combo2.Text <> "") And (Combo3.Text <> "") And (Combo4.Text <> "") And (Combo6.Text <> "") And _
(Combo8.Text <> "") And (Combo9.Text <> "") And (Combo10.Text <> "") And (Combo12.Text <> "") And _
(Combo1.Text <> "") And (Combo5.Text <> "") And (Combo7.Text <> "") And (Combo11.Text <> "") _
 Then
    UsuarioActual.Sexo = Combo1.Text
    FinishT = Timer
    TotalTimeT = FinishT - StartT
    UsuarioActual.TotalForma = UsuarioActual.TotalForma + TotalTimeT
    sBase = App.Path & "\Datos\Datos.mdb"
    Set cnn = New ADODB.Connection
    Set rst = New ADODB.Recordset
    cnn.Open "Provider=Microsoft.Jet.OLEDB.4.0; Data Source=" & sBase
    rst.Open "Select max(Numero) as num, count(*) as total from Usuario", cnn, adOpenDynamic, adLockOptimistic
    If rst.Fields!total = 0 Then numero = 0 Else numero = rst.Fields!num
    numero = numero + 1
    rst.Close
    rst.Open "SELECT * FROM Usuario", cnn, adOpenDynamic, adLockOptimistic
```

```
    With rst
      .AddNew
      '!Nombres = UsuarioActual.Nombres
      !numero = numero
      !Grupo = UsuarioActual.Grupo
      !Sexo = UsuarioActual.Sexo
     '  !Edad = UsuarioActual.Edad
      !Tiempo = UsuarioActual.TotalForma
      .Update
    End With
    ' Cerrar los objetos
    rst.Close
    cnn.Close
    Set rst = Nothing
    Set cnn = Nothing
    MsgBox "Usuario No. " & numero & Chr(13) & "Usted es parte del Grupo No.
" & UsuarioActual.Grupo & Chr(13) & "El tiempo total en el ingreso de sus datos
es:  " & UsuarioActual.TotalForma, vbOKOnly, "Experimento Multimodal"
    Unload Me
    Principal.Show
Else
    MsgBox "Por favor ingrese la información que se le solicita", vbCritical,
"Error"
End If
End Sub
Private Sub Command2_Click()
Unload Me
Principal.Show
End Sub
```

```
Private Sub Command3_Click()
StartT = Timer
Command3.Visible = Fals
Command1.Visible = True
Label1_1.Visible = True
Label1_2.Visible = True
Label1_3.Visible = True
Label1_4.Visible = True
Label1_5.Visible = True
Label1_6.Visible = True
Label1_7.Visible = True
Label1_8.Visible = True
Label1_9.Visible = True
Label1_10.Visible = True
Label1_11.Visible = True
Label1_12.Visible = True
Label2.Visible = True
Shape1.Visible = True
Label3.Visible = True
Shape2.Visible = True
Combo2.Visible = True
Combo3.Visible = True
Combo4.Visible = True
Combo6.Visible = True
Combo8.Visible = True
Combo9.Visible = True
Combo10.Visible = True
Combo12.Visible = True
Combo1.Visible = True
```

```
Combo5.Visible = True
Combo7.Visible = True
Combo11.Visible = True
End Sub
Private Sub Form_Load()
   UsuarioActual.TotalForma = 0
   Image1.Picture = LoadPicture(App.Path & "\imagenes\mouse.gif")
   Image2.Picture = LoadPicture(App.Path & "\imagenes\teclado.jpg")
   Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
   StartT = Timer
End Sub
```

**Test inicial para el entrenamiento del reconocimiento de voz.**

```
Option Explicit
Private Declare Function FindWindow Lib "user32" Alias _
   "FindWindowA" (ByVal lpClassName As String, ByVal lpWindowName _
   As String) As Long
Private Declare Function PostMessage Lib "user32" Alias "PostMessageA" _
   (ByVal hWnd As Long, ByVal wMsg As Long, ByVal wParam As Long, _
   ByVal lParam As Long) As Long
Private Const WM_COMMAND As Long = &H111
Private Const MIN_ALL As Long = 419
Private Const MIN_ALL_UNDO As Long = 416
Private Declare Function PlaySound Lib "winmm.dll" Alias "PlaySoundA" (ByVal
lpszName _
As String, ByVal hModule As Long, ByVal dwFlags As Long) As Long
Dim Voice As SpVoice
Const m_GrammarId = 12
```

```
Dim bSpeechInitialized As Boolean
Dim WithEvents RecoContext As SpSharedRecoContext
Dim Grammar As ISpeechRecoGrammar
Dim TopRule As ISpeechGrammarRule
Dim ActionTopRule As ISpeechGrammarRule
Private Declare Function ShellExecute Lib "shell32.dll" _
   Alias "ShellExecuteA" (ByVal hWnd As Long, ByVal _
   lpOperation As String, ByVal lpFile As String, ByVal _
   lpParameters As String, ByVal lpDirectory As String, _
   ByVal nShowCmd As Long) As Long
Const NumPalabras = 10
Dim Palabra(1 To NumPalabras) As String
Dim contador As Integer
Private Sub Command1_Click()
   If contador > 1 Then
      contador = contador - 1
   End If
   Label1.Caption = Palabra(contador)
   Image1.Visible = False
End Sub
Private Sub Command2_Click()
   Set RecoContext = Nothing
   Set Grammar = Nothing
   Set TopRule = Nothing
   Set ActionTopRule = Nothing
   Unload Me
   Instrucciones.Show
End Sub
Private Sub Command3_Click()
```

```
    contador = contador + 1
    If contador = NumPalabras + 1 Then
        Set RecoContext = Nothing
        Set Grammar = Nothing
        Set TopRule = Nothing
        Set ActionTopRule = Nothing
        Unload Me
        Instrucciones.Show
    Else
        Label1.Caption = Palabra(contador)
        Image1.Visible = False
    End If
End Sub
Private Sub Form_Load()
        Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
        Image1.Picture = LoadPicture(App.Path & "\imagenes\Ok.bmp")
        contador = 1
        Debug.Print "Initializing speech"
        Dim AfterCmdState As ISpeechGrammarRuleState
        Set RecoContext = New SpSharedRecoContext
        Set Grammar = RecoContext.CreateGrammar(m_GrammarId)
        Set TopRule = Grammar.Rules.Add("TopRule", SRATopLevel Or
SRADynamic, 1)
        Set ActionTopRule = Grammar.Rules.Add("ActionTopRule", SRATopLevel
Or SRADynamic, 2)
        'Set AfterCmdState = TopRule.AddState
        TopRule.InitialState.AddWordTransition Nothing, "Test"
        ActionTopRule.InitialState.AddWordTransition Nothing, "harley"
        Palabra(1) = "harley"
```

```
        ActionTopRule.InitialState.AddWordTransition Nothing, "pop"
        Palabra(2) = "pop"
        ActionTopRule.InitialState.AddWordTransition Nothing, "home"
        Palabra(3) = "home"
        ActionTopRule.InitialState.AddWordTransition Nothing, "the hammer"
        Palabra(4) = "the hammer"
        ActionTopRule.InitialState.AddWordTransition Nothing, "business"
        Palabra(5) = "business"
        ActionTopRule.InitialState.AddWordTransition Nothing, "software"
        Palabra(6) = "software"
        ActionTopRule.InitialState.AddWordTransition Nothing, "world"
        Palabra(7) = "world"
        ActionTopRule.InitialState.AddWordTransition Nothing, "card"
        Palabra(8) = "card"
        ActionTopRule.InitialState.AddWordTransition Nothing, "continental"
        Palabra(9) = "continental"
        ActionTopRule.InitialState.AddWordTransition Nothing, "finish"
        Palabra(10) = "finish"
        'RebuildGrammar
        Grammar.Rules.Commit
        Grammar.CmdSetRuleState "TopRule", SGDSActive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSInactive
        Set Voice = New SpVoice
        Set Voice.Voice = Voice.GetVoices().Item(0)
        Voice.Rate = 1
End Sub
Private Sub RecoContext_Hypothesis(ByVal StreamNumber As Long, ByVal
StreamPosition As Variant, ByVal _
Result As SpeechLib.ISpeechRecoResult)
```

```
    Debug.Print Result.PhraseInfo.GetText()
End Sub
Private Sub RecoContext_Recognition(ByVal StreamNumber As Long, ByVal
StreamPosition As Variant, _
ByVal RecognitionType As SpeechLib.SpeechRecognitionType, ByVal Result
As SpeechLib.ISpeechRecoResult)
    Dim s As String
    If Result.PhraseInfo.GetText() = "Test" Then
        Call PlaySound(App.Path & "\clicking.wav", 0, 0)
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        Label1.Caption = Palabra(contador)
        Command1.Visible = True
        Command3.Visible = True
    End If
    If Result.PhraseInfo.GetText() = Palabra(contador) Then
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        If Image1.Visible = True Then
            contador = contador + 1
            If contador = NumPalabras + 1 Then
                Set RecoContext = Nothing
                Set Grammar = Nothing
                Set TopRule = Nothing
                Set ActionTopRule = Nothing
                Unload Me
                Instrucciones.Show
            Else
                Label1.Caption = Palabra(contador)
```

```
            Image1.Visible = False
         End If
      Else
         Image1.Visible = True
      End If
   End If
End Sub
```

**Forma de trabajo del Grupo No. 2**

```
Option Explicit
Private Declare Function FindWindow Lib "user32" Alias _
   "FindWindowA" (ByVal lpClassName As String, ByVal lpWindowName _
   As String) As Long
Private Declare Function PostMessage Lib "user32" Alias "PostMessageA" _
   (ByVal hWnd As Long, ByVal wMsg As Long, ByVal wParam As Long, _
   ByVal lParam As Long) As Long
Private Const WM_COMMAND As Long = &H111
Private Const MIN_ALL As Long = 419
Private Const MIN_ALL_UNDO As Long = 416
Private Declare Function PlaySound Lib "winmm.dll" Alias "PlaySoundA" (ByVal
lpszName _
As String, ByVal hModule As Long, ByVal dwFlags As Long) As Long
Dim Voice As SpVoice
Const m_GrammarId = 12
Dim bSpeechInitialized As Boolean
Dim WithEvents RecoContext As SpSharedRecoContext
Dim Grammar As ISpeechRecoGrammar
Dim TopRule As ISpeechGrammarRule
```

```
Dim ActionTopRule As ISpeechGrammarRule
Dim MusicSelectRule As ISpeechGrammarRule
Private Type POINTAPI
   X As Long
   Y As Long
End Type
Private Declare Function ShellExecute Lib "shell32.dll" _
   Alias "ShellExecuteA" (ByVal hWnd As Long, ByVal _
   lpOperation As String, ByVal lpFile As String, ByVal _
   lpParameters As String, ByVal lpDirectory As String, _
   ByVal nShowCmd As Long) As Long
Private Declare Function GetCursorPos Lib "user32" (lpPoint As POINTAPI) As
Long
Private Declare Function ScreenToClient Lib "user32" (ByVal hWnd As Long,
lpPoint As POINTAPI) As Long
Dim Start, Finish, TotalTime, StartT, FinishT, TotalTimeT
Private sBase As String
Private cnn As ADODB.Connection
Private WithEvents rst As ADODB.Recordset
Function MouseX() As Long
   Dim lpPoint As POINTAPI
   GetCursorPos lpPoint
   MouseX = lpPoint.X
End Function
Function MouseY() As Long
   Dim lpPoint As POINTAPI
   GetCursorPos lpPoint
   MouseY = lpPoint.Y
End Function
```

```
Private Sub Form_Load()
    Image6.Picture = LoadPicture(App.Path & "\imagenes\voz.bmp")
    Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
    Debug.Print "Initializing speech"
    Dim AfterCmdState As ISpeechGrammarRuleState
    Set RecoContext = New SpSharedRecoContext
    Set Grammar = RecoContext.CreateGrammar(m_GrammarId)
    Set TopRule = Grammar.Rules.Add("TopRule", SRATopLevel Or
SRADynamic, 1)
    Set ActionTopRule = Grammar.Rules.Add("ActionTopRule", SRATopLevel
Or SRADynamic, 2)
    TopRule.InitialState.AddWordTransition Nothing, "start"
    ActionTopRule.InitialState.AddWordTransition Nothing, "male"
    ActionTopRule.InitialState.AddWordTransition Nothing, "female"
    ActionTopRule.InitialState.AddWordTransition Nothing, "The Hammer
Brothers"
    ActionTopRule.InitialState.AddWordTransition Nothing, "Blue Software"
    ActionTopRule.InitialState.AddWordTransition Nothing, "Business and
more"
    ActionTopRule.InitialState.AddWordTransition Nothing, "Sweet Home"
    ActionTopRule.InitialState.AddWordTransition Nothing, "World of Metal"
    ActionTopRule.InitialState.AddWordTransition Nothing, "administration
manager"
    ActionTopRule.InitialState.AddWordTransition Nothing, "engineer"
    ActionTopRule.InitialState.AddWordTransition Nothing, "financial manager"
    ActionTopRule.InitialState.AddWordTransition Nothing, "doctor"
    ActionTopRule.InitialState.AddWordTransition Nothing, "architect"
    ActionTopRule.InitialState.AddWordTransition Nothing, "Guatemala"
    ActionTopRule.InitialState.AddWordTransition Nothing, "Spain"
```

ActionTopRule.InitialState.AddWordTransition Nothing, "United Kingdom"
ActionTopRule.InitialState.AddWordTransition Nothing, "United States"
ActionTopRule.InitialState.AddWordTransition Nothing, "Germany"
ActionTopRule.InitialState.AddWordTransition Nothing, "Yes"
ActionTopRule.InitialState.AddWordTransition Nothing, "No"
ActionTopRule.InitialState.AddWordTransition Nothing, "Majestic Card"
ActionTopRule.InitialState.AddWordTransition Nothing, "Continental Gold"
ActionTopRule.InitialState.AddWordTransition Nothing, "Cash on line"
ActionTopRule.InitialState.AddWordTransition Nothing, "Money Card"
ActionTopRule.InitialState.AddWordTransition Nothing, "Price Card"
ActionTopRule.InitialState.AddWordTransition Nothing, "Finish"
ActionTopRule.InitialState.AddWordTransition Nothing, "Yahoo"
ActionTopRule.InitialState.AddWordTransition Nothing, "Hotmail"
ActionTopRule.InitialState.AddWordTransition Nothing, "Adidas"
ActionTopRule.InitialState.AddWordTransition Nothing, "Nike"
ActionTopRule.InitialState.AddWordTransition Nothing, "Kappa"
ActionTopRule.InitialState.AddWordTransition Nothing, "Reebok"
ActionTopRule.InitialState.AddWordTransition Nothing, "Mazda"
ActionTopRule.InitialState.AddWordTransition Nothing, "Mercedes Benz"
ActionTopRule.InitialState.AddWordTransition Nothing, "Seat"
ActionTopRule.InitialState.AddWordTransition Nothing, "Toyota"
ActionTopRule.InitialState.AddWordTransition Nothing, "Volvo"
ActionTopRule.InitialState.AddWordTransition Nothing, "Harley Davidson"
ActionTopRule.InitialState.AddWordTransition Nothing, "Honda"
ActionTopRule.InitialState.AddWordTransition Nothing, "Suzuki"
ActionTopRule.InitialState.AddWordTransition Nothing, "Yamaha"
ActionTopRule.InitialState.AddWordTransition Nothing, "Coca Cola"
ActionTopRule.InitialState.AddWordTransition Nothing, "Pepsi Cola"
ActionTopRule.InitialState.AddWordTransition Nothing, "dance"

```
        ActionTopRule.InitialState.AddWordTransition Nothing, "instrumental"
        ActionTopRule.InitialState.AddWordTransition Nothing, "pop"
        ActionTopRule.InitialState.AddWordTransition Nothing, "reggae"
        ActionTopRule.InitialState.AddWordTransition Nothing, "rock"
        ActionTopRule.InitialState.AddWordTransition Nothing, "techno"
        Grammar.Rules.Commit
        Grammar.CmdSetRuleState "TopRule", SGDSActive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSInactive
        Set Voice = New SpVoice
        Set Voice.Voice = Voice.GetVoices().Item(0)
        Voice.Rate = 1
End Sub
Private Sub RecoContext_Hypothesis(ByVal StreamNumber As Long, ByVal
StreamPosition As Variant, ByVal _
Result As SpeechLib.ISpeechRecoResult)
    Start = Timer
    Debug.Print Result.PhraseInfo.GetText()
End Sub
Private Sub RecoContext_Recognition(ByVal StreamNumber As Long, ByVal
StreamPosition As Variant, _
ByVal RecognitionType As SpeechLib.SpeechRecognitionType, ByVal Result
As SpeechLib.ISpeechRecoResult)
Dim s As String
Dim X, Y As Long
Dim numero As Integer
    X = MouseX
    Y = MouseY
    If Result.PhraseInfo.GetText() = "start" Then
        Call PlaySound(App.Path & "\clicking.wav", 0, 0)
```

```
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        LabelF1.Visible = True
        LabelF2.Visible = True
        LabelF3.Visible = True
        LabelF4.Visible = True
        LabelF5.Visible = True
        LabelF6.Visible = True
        LabelF7.Visible = True
        LabelF8.Visible = True
        LabelF9.Visible = True
        LabelF10.Visible = True
        LabelF11.Visible = True
        LabelF12.Visible = True
        Shape1.Visible = True
        Shape2.Visible = True
        Label2.Visible = True
        Label3.Visible = True
        StartT = Timer
    End If
    If Result.PhraseInfo.GetText() = "Finish" Then
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        If (LabelR1.Caption <> "") And (LabelR2.Caption <> "") And
(LabelR3.Caption <> "") And _
        (LabelR4.Caption <> "") And (LabelR5.Caption <> "") And
(LabelR6.Caption <> "") Then
            If LabelR1.Caption = "MALE" Then UsuarioActual.Sexo = "Masculino"
Else UsuarioActual.Sexo = "Femenino"
```

```
        FinishT = Timer
        TotalTimeT = FinishT - StartT
        UsuarioActual.TotalForma = UsuarioActual.TotalForma + TotalTimeT
        'MsgBox "Usted ha participado en el Grupo No. " & UsuarioActual.Grupo & Chr(13) & "El tiempo total en el ingreso de sus datos es:  " & UsuarioActual.TotalForma
        sBase = App.Path & "\Datos\Datos.mdb"
        Set cnn = New ADODB.Connection
        Set rst = New ADODB.Recordset
        cnn.Open "Provider=Microsoft.Jet.OLEDB.4.0; Data Source=" & sBase
        rst.Open "Select max(Numero) as num, count(*) as total from Usuario", cnn, adOpenDynamic, adLockOptimistic
        If rst.Fields!total = 0 Then numero = 0 Else numero = rst.Fields!num
        numero = numero + 1
        rst.Close
        rst.Open "SELECT * FROM Usuario", cnn, adOpenDynamic, adLockOptimistic
        With rst
           .AddNew
              '!Nombres = UsuarioActual.Nombres
              !numero = numero
              !Grupo = UsuarioActual.Grupo
              !Sexo = UsuarioActual.Sexo
              '!Edad = UsuarioActual.Edad
              !Tiempo = UsuarioActual.TotalForma
           .Update
        End With
        rst.Close
        cnn.Close
```

```
            Set rst = Nothing
            Set cnn = Nothing
            Set RecoContext = Nothing
            Set Grammar = Nothing
            Set TopRule = Nothing
            Set ActionTopRule = Nothing
            MsgBox "Usuario No. " & numero & Chr(13) & "Usted es parte del Grupo No. " & UsuarioActual.Grupo & Chr(13) & "El tiempo total en el ingreso de sus datos es:  " & UsuarioActual.TotalForma, vbOKOnly, "Experimento Multimodal"
            Unload Me
            Principal.Show
        Else
            MsgBox "Por favor ingrese la información que se le solicita", vbCritical, "Error"
        End If
    End If
    If (X >= 80 And X <= 247) And (Y >= 188 And Y <= 211) Then
     If (Result.PhraseInfo.GetText() = "male") Or (Result.PhraseInfo.GetText() = "female") Then
            LabelR1.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line2.Visible = True
      End If
    End If
    If (X >= 80 And X <= 247) And (Y >= 267 And Y <= 290) Then
     If (Result.PhraseInfo.GetText() = "The Hammer Brothers") Or (Result.PhraseInfo.GetText() = "Blue Software") Or _
```

```
        (Result.PhraseInfo.GetText() = "Business and more") Or
(Result.PhraseInfo.GetText() = "Sweet Home") Or _
        (Result.PhraseInfo.GetText() = "World of Metal") Then
          LabelR2.Caption = UCase(Result.PhraseInfo.GetText())
          Grammar.CmdSetRuleState "TopRule", SGDSInactive
          Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
          Line3.Visible = True
      End If
    End If
    If (X >= 80 And X <= 247) And (Y >= 347 And Y <= 370) Then
      If (Result.PhraseInfo.GetText() = "administration manager") Or
(Result.PhraseInfo.GetText() = "architect") Or _
        (Result.PhraseInfo.GetText() = "engineer") Or
(Result.PhraseInfo.GetText() = "doctor") Or _
        (Result.PhraseInfo.GetText() = "financial manager") Then
          LabelR3.Caption = UCase(Result.PhraseInfo.GetText())
          Grammar.CmdSetRuleState "TopRule", SGDSInactive
          Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
          Line4.Visible = True
      End If
    End If
    If (X >= 80 And X <= 247) And (Y >= 427 And Y <= 450) Then
      If (Result.PhraseInfo.GetText() = "Germany") Or
(Result.PhraseInfo.GetText() = "Guatemala") Or _
        (Result.PhraseInfo.GetText() = "Spain") Or (Result.PhraseInfo.GetText() =
"United Kingdom") Or _
        (Result.PhraseInfo.GetText() = "United States") Then
          LabelR4.Caption = UCase(Result.PhraseInfo.GetText())
          Grammar.CmdSetRuleState "TopRule", SGDSInactive
```

```
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        Line5.Visible = True
     End If
   End If
   If (X >= 80 And X <= 247) And (Y >= 507 And Y <= 546) Then
     If (Result.PhraseInfo.GetText() = "Yes") Or (Result.PhraseInfo.GetText() =
"No") Then
        LabelR5.Caption = UCase(Result.PhraseInfo.GetText())
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        Line6.Visible = True
      End If
   End If
   If (X >= 80 And X <= 247) And (Y >= 603 And Y <= 642) Then
     If (Result.PhraseInfo.GetText() = "Majestic Card") Or
(Result.PhraseInfo.GetText() = "Price Card") Or _
       (Result.PhraseInfo.GetText() = "Continental Gold") Or
(Result.PhraseInfo.GetText() = "Cash on line") Or _
       (Result.PhraseInfo.GetText() = "Money Card") Then
        LabelR6.Caption = UCase(Result.PhraseInfo.GetText())
        Grammar.CmdSetRuleState "TopRule", SGDSInactive
        Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
        Line7.Visible = True
     End If
   End If
   If (X >= 528 And X <= 695) And (Y >= 188 And Y <= 211) Then
     If (Result.PhraseInfo.GetText() = "Hotmail") Or (Result.PhraseInfo.GetText()
= "Yahoo") Then
        LabelR7.Caption = UCase(Result.PhraseInfo.GetText())
```

```
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line8.Visible = True
        End If
    End If
    If (X >= 528 And X <= 695) And (Y >= 267 And Y <= 290) Then
        If (Result.PhraseInfo.GetText() = "Adidas") Or (Result.PhraseInfo.GetText()
= "Nike") Or _
            (Result.PhraseInfo.GetText() = "Kappa") Or (Result.PhraseInfo.GetText()
= "Reebok") Then
            LabelR8.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line9.Visible = True
        End If
    End If
    If (X >= 528 And X <= 695) And (Y >= 347 And Y <= 370) Then
        If (Result.PhraseInfo.GetText() = "Mazda") Or (Result.PhraseInfo.GetText()
= "Toyota") Or _
            (Result.PhraseInfo.GetText() = "Mercedes Benz") Or
(Result.PhraseInfo.GetText() = "Seat") Or _
            (Result.PhraseInfo.GetText() = "Volvo") Then
            LabelR9.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line10.Visible = True
        End If
    End If
    If (X >= 528 And X <= 695) And (Y >= 427 And Y <= 450) Then
```

```
    If (Result.PhraseInfo.GetText() = "Harley Davidson") Or
(Result.PhraseInfo.GetText() = "Suzuki") Or _
        (Result.PhraseInfo.GetText() = "Yamaha") Or (Result.PhraseInfo.GetText()
= "Honda") Then
            LabelR10.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line11.Visible = True
     End If
   End If
   If (X >= 528 And X <= 695) And (Y >= 507 And Y <= 546) Then
     If (Result.PhraseInfo.GetText() = "Coca Cola") Or
(Result.PhraseInfo.GetText() = "Pepsi Cola") Then
            LabelR11.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
            Line12.Visible = True
     End If
   End If
   If (X >= 528 And X <= 695) And (Y >= 603 And Y <= 642) Then
     If (Result.PhraseInfo.GetText() = "dance") Or (Result.PhraseInfo.GetText() =
"rock") Or _
        (Result.PhraseInfo.GetText() = "instrumental") Or
(Result.PhraseInfo.GetText() = "reggae") Or _
        (Result.PhraseInfo.GetText() = "techno") Or (Result.PhraseInfo.GetText()
= "pop") Then
            LabelR12.Caption = UCase(Result.PhraseInfo.GetText())
            Grammar.CmdSetRuleState "TopRule", SGDSInactive
            Grammar.CmdSetRuleState "ActionTopRule", SGDSActive
```

```
            Line13.Visible = True
        End If
    End If
End Sub
```

**Detalle del Grupo de Trabajo, según usuario.**

```
Private Sub Form_Load()
Dim intRecord As Integer
Dim intField As Integer
    With Adodc1
        .ConnectionString = "Provider=Microsoft.Jet.OLEDB.4.0; Data Source = " &
App.Path & "\Datos\Datos.mdb"
        .RecordSource = "Select numero as Usuario, sexo, tiempo from Usuario
where grupo = " & Str(EstadisticasGrupo) & " order by numero"
        .Refresh
    End With

    Set dgGrupo.DataSource = Adodc1
    For i = 0 To 2
        Set Text1(i).DataSource = Adodc1
    Next
    Text1(0).DataField = "Usuario"
    Text1(1).DataField = "Sexo"
    Text1(2).DataField = "Tiempo"
  intRecord = Adodc1.Recordset.RecordCount
  intField = Adodc1.Recordset.Fields.Count
  Label3.Caption = intRecord
  If EstadisticasGrupo = 1 Then
```

```
    Image1.Visible = True
    Image2.Visible = True
    Image2.Picture = LoadPicture(App.Path & "\imagenes\mouse.gif")
    Image1.Picture = LoadPicture(App.Path & "\imagenes\teclado.jpg")
    LabelAccion.Caption = "Teclado y ratón"
    LabelGrupo.Caption = "GRUPO 1"
  Else
    Image3.Visible = True
    Image3.Picture = LoadPicture(App.Path & "\imagenes\voz.bmp")
    LabelAccion.Caption = "Apuntar y hablar"
    LabelGrupo.Caption = "GRUPO 2"
  End If
  Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
End Sub
```

**Acceso a los resultados globales del experimento.**

```
Private WithEvents rst As ADODB.Recordset
Private Sub Command1_Click()
EstadisticasGrupo = 1
DetalleGrupo.Show
End Sub
Private Sub Command2_Click()
EstadisticasGrupo = 2
DetalleGrupo.Show
End Sub
Private Sub Command3_Click()
    Unload Me
    Principal.Show
```

```
End Sub
Private Sub Form_Load()
    WebBrowser1.Navigate App.Path & "\imagenes\logo.gif"
    Me.Picture = LoadPicture(App.Path & "\imagenes\fondo.jpg")
    sBase = App.Path & "\Datos\Datos.mdb"
    Set cnn = New ADODB.Connection
    Set rst = New ADODB.Recordset
    cnn.Open "Provider=Microsoft.Jet.OLEDB.4.0; Data Source=" & sBase
    rst.Open "SELECT (sum(tiempo)/Count(*)) as prom, count(*) as num From
Usuario WHERE grupo=1", cnn, adOpenDynamic, adLockOptimistic
    PromGuno = rst.Fields!Prom
    NumGuno = rst.Fields!num
    rst.Close
    rst.Open "SELECT (sum(tiempo)/Count(*)) as prom, count(*) as num From
Usuario WHERE grupo=2", cnn, adOpenDynamic, adLockOptimistic
    PromGdos = rst.Fields!Prom
    NumGdos = rst.Fields!num
    rst.Close
    rst.Open "SELECT count(*) as numero, (sum(tiempo)/Count(*)) as promh
From Usuario WHERE grupo=1 and sexo=" & Chr(34) & "Masculino" & Chr(34),
cnn, adOpenDynamic, adLockOptimistic
    male1 = rst.Fields!numero
    If rst.Fields!numero = 0 Then
       malep1 = 0
    Else
       malep1 = rst.Fields!promh
    End If
    rst.Close
```

```
    rst.Open "SELECT count(*) as numero, (sum(tiempo)/Count(*)) as promm From Usuario WHERE grupo=1 and sexo=" & Chr(34) & "Femenino" & Chr(34), cnn, adOpenDynamic, adLockOptimistic
    female1 = rst.Fields!numero
    If rst.Fields!numero = 0 Then
        femalep1 = 0
    Else
        femalep1 = rst.Fields!promm
    End If
    rst.Close
    LabelNp1.Caption = NumGuno
    LabelNh1.Caption = male1
    LabelNm1.Caption = female1
    LabelTph1.Caption = Format(malep1, "###0.000")
    LabelTpm1.Caption = Format(femalep1, "###0.000")
    LabelTp1.Caption = Format(PromGuno, "###0.000")
    rst.Open "SELECT count(*) as numero, (sum(tiempo)/Count(*)) as promh From Usuario WHERE grupo=2 and sexo=" & Chr(34) & "Masculino" & Chr(34), cnn, adOpenDynamic, adLockOptimistic
    male2 = rst.Fields!numero
    If rst.Fields!numero = 0 Then
        malep2 = 0
    Else
        malep2 = rst.Fields!promh
    End If
    rst.Close
    rst.Open "SELECT count(*) as numero, (sum(tiempo)/Count(*)) as promm From Usuario WHERE grupo=2 and sexo=" & Chr(34) & "Femenino" & Chr(34), cnn, adOpenDynamic, adLockOptimistic
```

```
    female2 = rst.Fields!numero
    If rst.Fields!numero = 0 Then
       femalep2 = 0
    Else
       femalep2 = rst.Fields!promm
    End If
    rst.Close
    LabelNp2.Caption = NumGdos
    LabelNh2.Caption = male2
    LabelNm2.Caption = female2
    LabelTph2.Caption = Format(malep2, "###0.000")
    LabelTpm2.Caption = Format(femalep2, "###0.000")
    LabelTp2.Caption = Format(PromGdos, "###0.000")
    LabelNmt.Caption = female1 + female2
    LabelNht.Caption = male1 + male2
    LabelNpt.Caption = NumGuno + NumGdos
    LabelTpmt.Caption = Format(((femalep1 + femalep2) / 2), "###0.000")
    LabelTpht.Caption = Format(((malep1 + malep2) / 2), "###0.000")
    LabelTpt.Caption = Format(((PromGuno + PromGdos) / 2), "###0.000")
Totalf = femalep1 + femalep2
    If Totalf = 0 Then
       PorFemaleUno = 0
       PorFemaleDos = 0
    Else
       PorFemaleUno = (femalep1 * 100) / Totalf
       PorFemaleDos = (femalep2 * 100) / Totalf
    End If
    Totalm = malep1 + malep2
    If Totalm = 0 Then
```

```
        PorMaleUno = 0
        PorMaleDos = 0
    Else
        PorMaleUno = (malep1 * 100) / Totalm
    PorMaleDos = (malep2 * 100) / Totalm
    End If
    total = PromGuno + PromGdos
    PorcentajeUno = (PromGuno * 100) / total
    PorcentajeDos = (PromGdos * 100) / total
    Dim pie3 As New clsPie
    Dim pie1 As New clsPie
    Dim pie2 As New clsPie
    Labelazul3.Caption = CInt(PorcentajeUno) & " % del tiempo total"
    Labelrojo3.Caption = CInt(PorcentajeDos) & " % del tiempo total"
    pie3.AddSection CInt(PorcentajeUno), vbBlue
    pie3.AddSection CInt(PorcentajeDos), vbRed
    pie3.DrawPie pic3, 75
    Labelazul1.Caption = CInt(PorFemaleUno) & " % del tiempo total"
    Labelrojo1.Caption = CInt(PorFemaleDos) & " % del tiempo total"
    pie1.AddSection CInt(PorFemaleUno), vbBlue
    pie1.AddSection CInt(PorFemaleDos), vbRed
    pie1.DrawPie pic1, 75
    Labelazul2.Caption = CInt(PorMaleUno) & " % del tiempo total"
    Labelrojo2.Caption = CInt(PorMaleDos) & " % del tiempo total"
    pie2.AddSection CInt(PorMaleUno), vbBlue
    pie2.AddSection CInt(PorMaleDos), vbRed
    pie2.DrawPie pic2, 75
End Sub
```

# ANEXO

# SDK SAPI 5.1

## Sobre el SDK

El "Microsoft® Speech SDK" ha sido diseñado para trabajar con la interfaz de lenguaje (SAPI), con lo cual, Microsoft® da continuidad a los motores de reconocimiento de habla proporcionando herramientas para reconocimiento de texto-habla. Microsoft® SDK incluye herramientas, ejemplos y documentación por construir aplicaciones del habla.

## SAPI 5.1

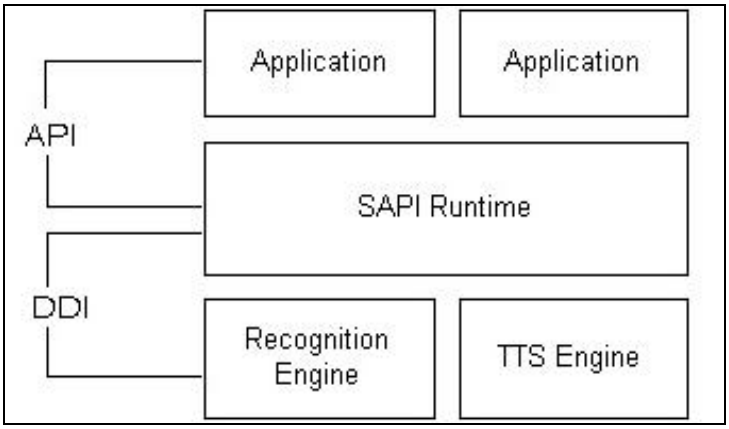
La interfaz de aplicación (API) SAPI, reduce dramáticamente el código requerido para utilizar una aplicación de reconocimiento del habla y de texto a voz, creando una tecnología mucho más accesible y robusta para un amplio rango de aplicaciones. Para ampliar sobre este tema, se definen los siguientes aspectos a destacar:

- Descripción global del API
- API para el texto a voz
- API para reconocimiento de voz

**Descripción global de API**

El SAPI proporciona una interfaz de alto nivel entre una aplicación y las herramientas de voz. SAPI implementa y cuida hasta el más mínimo detalle que se necesita para controlar y manejar el funcionamiento, en tiempo real, de varias herramientas para capturar la voz. Los dos tipos básicos de herramientas de SAPI son texto a voz TTS (por sus siglas en inglés "*text to speech*") los sistemas y reconocedores de voz. Los sistemas de TTS sintetiza y unifica el texto y los archivos en audio de la voz utilizando voces sintéticas. Los reconocedores del habla se convierten en una voz humana del texto leído, unificando textos y archivos de voz.

**Figura 22. Estructura de SAPI**




**API para el texto a voz**

Las aplicaciones pueden controlar texto a voz (TTS) usando la interfaz del *ISpVoice COM* (*Component Object Model* por sus siglas en inglés). Cuando una aplicación ha creado un objeto de *ISpVoice*, la aplicación únicamente necesita llamar a *ISpVoice::Speak* para generar una salida de voz desde algún texto. Además, la interfaz de *IspVoice* también mantiene varios métodos de voz

para cambiar las propiedades de la velocidad *con ISpVoice::SetRate*, el volumen de salida con *ISpVoice::SetVolume* y cambiar la voz de salida actual con *ISpVoice::SetVoice*

También se puede insertar comandos especiales de SAPI con el texto de la entrada, para cambiar propiedades de síntesis en tiempo real como la voz, tono, énfasis en palabras, velocidad y volumen del hablado. Este valor agregado de la síntesis sapi.xsd, que utiliza una estructura normal XML, es una manera simple pero poderosa de personalizar el discurso de TTS, independiente del motor específico o voz que se encuentre actualmente en uso.

El método de *IspVoice::Speak* puede operar de forma síncronamente (retorna únicamente cuando ha terminado de hablar) o asincrónicamente (retorna inmediatamente y habla como un proceso en segundo plano o "*backgruound*"). Al hablar asincrónicamente (*SPF_ASYNC*), la información de estado de tiempo real como el estado del habla y la colocación del texto actualmente analizado se puede registrar utilizando *ISpVoice::GetStatus*. También mientras se habla asincrónicamente, el nuevo texto puede ser hablado por cualquiera e interrumpirá inmediatamente la salida actual de voz (*SPF_PURGEBEFORESPEAK*), o automáticamente añadiendo el nuevo texto al final de la salida actual. Además de la interfaz de *ISpVoice*, SAPI proporciona también muchas utilidades con interfaces COM para aplicaciones TTS más avanzadas.

**Eventos**

SAPI se comunica con aplicaciones enviando eventos que usan mecanismos estándar de retorno de llamada o "*callback*" (Mensaje de Ventana, proceso "*callback*" o evento de Win32). Para TTS, se usan eventos

principalmente para sincronizar la salida del habla. Las aplicaciones pueden sincronizarse con acciones en tiempo real y cuando ocurran procesos como palabras mal pronunciadas, fonemas (animación de la boca) que limiten el reconocimiento o marcadores personalizados de la aplicación. Las aplicaciones pueden inicializar y manejar estos eventos en tiempo real utilizando *ISpNotifySource, ISpNotifySink, ISpNotifyTranslator, ISpEventSink, ISpEventSource, y ISpNotifyCallback*.

**Componentes Léxicos**

Las aplicaciones pueden mantener pronunciaciones de palabra personalizadas para los motores de síntesis de voz, que utilizan métodos proporcionados por *ISpContainerLexicon, ISpLexicon y ISpPhoneConverter*.

**Recursos**

Encontrando y seleccionando SAPI como archivos de voz y pronunciación, obtenidos del análisis del habla, se pueden utilzar las siguientes interfaces COM: *ISpDataKey, ISpRegDataKey, ISpObjectTokenInit, ISpObjectTokenCategory, ISpObjectToken, IEnumSpObjectTokens, ISpObjectWithToken, ISpResourceManager y ISpTask.*

**Audio**

Finalmente, hay una interfaz para personalizar la salida del audio a algún destino especial como telefonía y personalización de hardware (*ISpAudio, ISpMMSysAudio, ISpStream, ISpStreamFormat, ISpStreamFormatConverter*).

**API para el Reconocimiento del habla.**

Así como *ISpVoic*e es la interfaz principal para la síntesis del habla, *ISpRecoContext* es la interfaz principal para el reconocimiento de voz. Como el *ISpVoice*, el *ISpRecoContext* maneja eventos que se utilizan como medio de transporte en la aplicación de voz, para las notificaciones de los eventos receptores de reconocimiento del habla solicitados.

Una aplicación tiene la opción de dos tipos diferentes de motores de reconocimiento del habla (*ISpRecognizer*). Lo recomendable para la mayoría de las aplicaciones del habla, es contar con un reconocedor compartido que posiblemente, podrá interactuar con otras aplicaciones de reconocimiento de voz. Para crear un *ISpRecoContext* para un *ISpRecognizer* compartido, una aplicación sólo necesita llamar a *CoCreateInstance* del componente COM *CLSID_SpSharedRecoContext*. En este caso, SAPI preparará el flujo de entrada de audio, configurado con los valores por defecto de SAPI. Para grandes aplicaciones de servidor, el reconocedor se ejecutará exclusivamente en forma aislada en el sistema, siendo ésto clave para que el desempeño sea importante en los procesos del motor de reconocimiento del habla.

El orden para crear un *ISpRecoContext* para un *InProc ISpRecognizer*, es de la siguiente forma: en primer lugar, la aplicación debe llamar al *CoCreateInstance* en el componente *CLSID_SpInprocRecoInstance* para crear su propio *InProc ISpRecognizer*. Entonces, la aplicación debe hacer una llamada a *ISpRecognizer::SetInput* para preparar la entrada del audio. Finalmente, la aplicación llama a *ISpRecognizer::CreateRecoContext* para obtener un *ISpRecoContext*.

El próximo paso, es colocar notificaciones o procesos para los eventos que la aplicación está interesada en reconocer. Cuando el *ISpRecognizer* también es un *ISpEventSource* que a su vez es un *ISpNotifySource*, la aplicación puede llamar uno de los métodos de *ISpNotifySource* de su *ISpRecoContext* e indicar dónde deben informarse los eventos para ese *ISpRecoContext*. En ese momento, debe llamar *ISpEventSource::SetInterest* para indicar qué eventos necesita notificar. El evento más importante es el *SPEI_RECOGNITION* que indica que el *ISpRecognizer* ha reconocido alguna palabra para este *ISpRecoContext*.

Finalmente, una aplicación del habla debe crear, cargar y activar un *ISpRecoGrammar,* que esencialmente, indica qué tipo de pronunciaciones debe reconocer, por ejemplo, un dictado o un control de la gramática. Primero, la aplicación crea un *ISpRecoGrammar* usando *ISpRecoContext::CreateGrammar*. Entonces, la aplicación carga la gramática apropiada, realizando la llamada a *ISpRecoGrammar::LoadDictation* para dictado o uno de los métodos de *ISpRecoGrammar::LoadCmdxxx* para el control generado en la gramática. Como últmo paso, para activar estas gramáticas de reconocimiento, la aplicación llama a *ISpRecoGrammar::SetDictationState* para dictado o *ISpRecoGrammar::SetRuleState* o *ISpRecoGrammar::SetRuleIdState* para el control de gramática.

Cuando los reconocimientos regresan a la aplicación por medio del mecanismo de la notificación solicitada, el miembro del *lParam* de la estructura de *SPEVENT* será un *ISpRecoResult,* por lo que la aplicación puede determinar lo que se reconoció y para cual de los *ISpRecoGrammar* del *ISpRecoContext* se ha reconocido.

Un *ISpRecognizer*, compartido o no, puede tener *ISpRecoContexts* múltiples asociado con él, y cada uno puede notificarse de su propia manera de eventos que pertenecen a él. De igual forma, un *ISpRecoContext* puede tener *ISpRecoGrammars* múltiples asociados con él, cada uno por reconocer tipos diferentes de pronunciaciones.