Capítulo 1.

Reconocimiento de Voz y Educación

El estado del arte en sistemas con Reconocimiento de Voz ha progresado impresionantemente. Una interfaz bien diseñada puede tomar las ventajas que ofrece un reconocedor y compensar sus áreas débiles, también puede implementar los principios básicos para el aprendizaje de un segundo idioma, con lo cual se tienen todos los componentes necesarios de un instructor del lenguaje [Eskenazi, 99].

1.1 La Tecnología del Habla

La Tecnología del Habla se estructura en cuatro tecnologías básicas principales:

El Reconocimiento de Voz o Reconocimiento del Habla. Es el proceso de conversión de un mensaje hablado en texto, que permite al usuario una comunicación con la computadora.

La Síntesis de Voz o Conversión Texto a Voz. Se ocupa de la generación de mensajes hablados mediante la simulación del proceso de lectura de un texto escrito almacenado en formato electrónico.

El Reconocimiento de Locutores. Es el proceso de identificación o verificación de la identidad del hablante de forma automática a partir de la señal de voz.

La Codificación de Voz. Su objetivo es la búsqueda de representaciones eficientes en formato digital de la señal de voz para su almacenamiento y/o transmisión, persiguiendo obtener la mayor calidad posible, para el menor número de bits por muestra.

Podríamos, por tanto, situar a la Tecnología del Habla como receptora de un amplio conjunto de conocimientos y procedimientos de actuación sobre la información representada en la señal de voz. Conocimientos que se articulan con un alto grado de dificultad y especialización, ya que pertenecen a un marco científico-técnico multidisciplinario, donde se dan cita diferentes ramas del saber como son: fisiología, acústica, lingüística, procesado digital de señales, inteligencia artificial, teoría de la comunicación y de la información, y ciencias de la computación.

1.1.1 Reconocimiento automático de voz

El Reconocimiento Automático de Voz, como se mencionó anteriormente, es el proceso de convertir palabras habladas, capturadas por un micrófono o teléfono, en un conjunto de palabras escritas.

Las principales características que diferencian a los sistemas basados en Reconocimiento de Voz frente a otras alternativas son: la naturalidad que supone utilizar el habla en las operaciones de comando y control, y la precisión y robustez en la comunicación para diferentes usuarios y diferentes entornos. El estado actual de la investigación en Reconocimiento del Voz nos muestra excelentes resultados de sistemas trabajando en entornos controlados de laboratorio. Sin embargo, una aplicación real de esta tecnología exige un funcionamiento en el mundo real donde el grado de dificultad de los problemas es un orden de magnitud mayor.

Los sistemas de Reconocimiento de Voz se caracterizan por muchos parámetros, algunos de ellos se muestran en la siguiente tabla:

Tabla 1.1 Parámetros que caracterizan a un Sistema de reconocimiento de voz

Parámetros	Rango
Modo de hablar	Palabras aisladas o habla continua
Estilo del habla	Voz de lectura o voz espontánea
Aislamiento	Dependiente del locutor o Independiente del locutor
Vocabulario	Pequeño (<20 palabras) o grande (>20,000 palabras)
Modelo del lenguaje	Estados finitos o dependiente del contexto
Perplejidad	Pequeña (<10) o larga (>100)
Reducción del ruido en el habla	Alta (>30 dB) o baja (<10 dB)

El reconocimiento de la voz es un problema difícil, debido a muchas fuentes de variabilidad asociadas con la señal acústica, cambios en el ambiente, cambios en el estado físico o emocional del locutor, o el tamaño del tracto bocal.

1.1.2 La Evolución de los Sistemas de Reconocimiento Automático de Voz

Las primeras investigaciones en el desarrollo de éstos sistemas fueron realizadas en la década de los 50's. Los estudios trataron de explotar las ideas fundamentales de la fonética acústica.

Durante la década de los 60's los estudios se enfocaron,

principalmente, a los problemas de segmentación, clasificación y reconocimiento de patrones.

En los 70's se mejoró la tecnología de reconocimiento para palabras aisladas y continuas. Se hicieron reconocedores que aceptaban un vocabulario más extenso. También se desarrollaron técnicas como: time warping, modelado probabilístico y el algoritmo de retropropagación [Rabiner & Juang, 93; Rumelhart & McClelland, 86].

En la década de los 80's hubo un cambio en la tecnología, del enfoque basado en reconocimiento de patrones a métodos de modelado probabilístico, como el método de cadenas ocultas de Markov (HMM) [Rabiner & Juang, 93]. Las redes neuronales se reintrodujeron para resolver problemas de reconocimiento de voz [Waibel & Lee, 90].

En la actualidad, existen diversos factores que contribuyen al mejoramiento y el progreso de los sistemas de Reconocimiento de Voz, como los HMM y las redes neuronales. Se han realizado grandes esfuerzos para desarrollar una base de datos de voz con un vocabulario grande, el cual pueda ser usado en el entrenamiento, desarrollo y prueba de los estos sistemas. Por otra parte, el establecimiento de estándares para la evaluación del desempeño en el reconocimiento permite hacer comparaciones entre distintos sistemas. Gracias a los avances en la tecnología computacional, los sistemas pueden ser probados en tiempo real sin la necesidad de hardware adicional [Cole et al., 99].

La evolución de éstos tiene como algunos representantes a los siguientes sistemas [Hernández et al., 94]:

ATR HMM-LR. Sistema japonés desarrollado en ATR. Está basado en procedimientos específicos de modelado de sonidos que no utilizan estructuras intermedias de modelos de fonema o palabra.

AT&T y BELL NORTHERN RESEARCH. Ambos Sistemas incorporan procedimientos específicos para aplicaciones de automatización de servicios telefónicos.

BYBLOS. Desarrollado por BBN. Byblos es el nombre de una ciudad fenicia donde se descubrió la primera muestra de escritura fonética. Este detalle marca el énfasis que se pone actualmente en desarrollar Sistemas sobre una base fonética. Aunque se trata de un sistema dependiente del locutor, este sistema ha aportado un nuevo y eficiente procedimiento de reconocimiento rápido (búsqueda rápida) basado en algoritmos N-best.

CSELT. Desarrollado en el centro italiano del mismo nombre. Su principal innovación es un sistema de búsqueda rápida basada en un primer descifrado fonético simple y rápido, seguido por una búsqueda más detallada.

DECIPHER. Desarrollado en SRI International. Su principal novedad fue la representación detallada de aspectos fonéticos importantes, tales

como la coarticulación entre palabras.

LINCOLN. Desarrollado en el laboratorio del mismo nombre. Su principal aportación es el modelado de voz rápida, con emoción, tensión, etc.

PHILIPS. Desarrollado por la empresa del mismo nombre. Es un sistema pionero en procesos de reconocimiento rápidos para habla continua y vocabularios de hasta 10,000 palabras.

SPHINX-II. Desarrollado en la Universidad de Carnegie-Mellon. Es un sistema pionero en reconocimiento independiente de locutor para grandes vocabularios.

TANGORA. Desarrollado en IBM. Se trata de un sistema dependiente del locutor para grandes vocabularios. Su principal interés es un proceso de adaptación a un nuevo locutor que requiere 20 minutos para leer 100 frases de 1200 palabras, 700 de las cuales son distintas.

En los últimos diez años se ha producido un notable avance que hace posible disponer de una tecnología básica capaz de soportar aplicaciones y servicios comerciales. En Reconocimiento de Voz, se han conseguido reconocedores que, aunque limitados en cuanto al tamaño del vocabulario, poseen una calidad suficiente para soportar un gran número de aplicaciones. Como productos comerciales, están disponibles reconocedores de dígitos aislados y concatenados, y reconocedores de palabras aisladas que manejan vocabularios de miles de palabras y, lo que es más importante, es posible definir el vocabulario del reconocedor sin necesidad de realizar un largo y costoso proceso de entrenamiento (reconocedor de vocabulario libre). En fase precompetitiva (prototipos de laboratorio) existen reconocedores de habla continua capaces de manejar vocabularios de algunos miles de palabras.

El desarrollo de un componente de diálogo y los aspectos de un modelo de interacción en sistemas interactivos de voz se encuentran soportados actualmente en términos de herramientas y técnicas avanzadas como lo son [Ole et al., 98]:

Wizard of Oz. Es un método experimental prototipo en el cual un humano (el mago) simula una parte o todo el modelo interactivo del sistema a ser desarrollado y puede ponerse en interacción con los usuarios los cuales creerán que están interactuando con un sistema real.

Manejo del Corpus. Existen muchas herramientas de éste tipo, el Text Encoding Initiative (TEI) es la herramienta más amplia que existe para la representación de texto incluyendo transcripción del habla.

Modelo de implementación del dialogo. Existen muchas herramientas de éste tipo, DDLTool es un editor gráfico que soporta la representación de un software de manejo del diálogo en el Lenguaje Descriptor de Diálogo. CSLUrp es un ambiente de gráfico de desarrollo

de prototipos que es muy similar al DDLTool en muchos aspectos.

Experimentación y desarrollo. Existen muchas técnicas y herramientas para desarrollar y experimentar, como DDLTool, parte de CSLUrp, Gnu's C++.

Evaluación. En la colaboración DARPA ATIS una herramienta de software fue desarrollada para comparar automáticamente un conjunto de respuestas con aquellas producidas por varios sistemas.

Toolkits. El Oregon Graduate Institute (OGI) hizo recientemente un toolkit (caja de herramientas), disponible en el web, llamado CSLU Toolkit.

Existen sistemas comerciales desarrollados por compañías como AT&T, SpeechWorks, Dragon Systems y otras, los cuales han tenido gran aceptación pues desarrollan aplicaciones para el mundo real como bancos, finanzas, seguros, agencias de viajes, tiempos compartidos, entre otros [Blyth & Piper, 94].

1.1.3 Arquitectura de un sistema de un reconocedor automático de voz

Los principales componentes de un sistema de reconocimiento de voz se muestran en la siguiente figura.

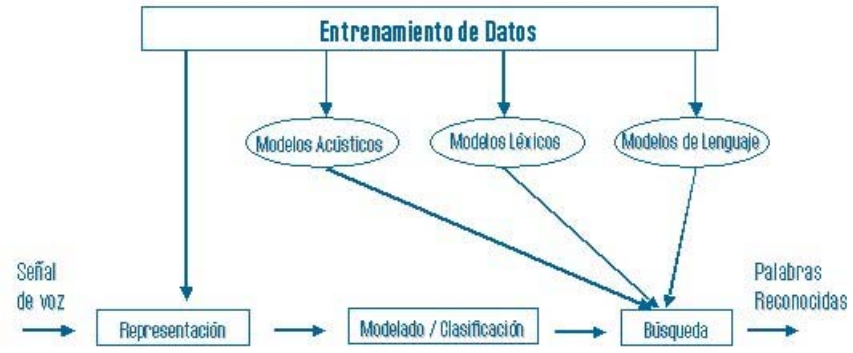

Figura 1.1 Componentes de un sistema típico de reconocimiento de voz.

La señal de voz digitalizada se transforma en un conjunto de medidas útiles o características de manera fija, típicamente una cada 10-20 ms. Estas características son usadas para buscar la palabra con mayor probabilidad, haciendo uso de restricciones impuestas por modelos acústicos, léxicos y del lenguaje.

1.1.4 El proceso de reconocimiento automático de voz

El proceso de reconocimiento automático de voz consiste en:

1. Obtener y digitalizar la señal de voz
2. Extraer un conjunto de características esenciales de la señal
3. Introducir las características a un clasificador
4. Realizar un algoritmo de búsqueda para encontrar la secuencia permitida más probable utilizando la salida obtenida y una red de pronunciaciones.
5. Encontrar la(s) palabra(s) que se desea reconocer.

En la etapa de extracción de características la señal de voz se divide en una colección de segmentos. Luego, se obtiene una representación de las características acústicas más significativas para cada segmento, esto se hace aplicando alguna técnica de procesamiento de señales. Con dichas características se construye un conjunto de vectores que constituyen la entrada al clasificador.

El clasificador aplica un modelo probabilístico y vincula a cada uno de los vectores de características con alguna unidad lingüística (palabra, fonema u otra unidad específica). Posteriormente se realiza la búsqueda para encontrar la secuencia de segmentos con mayor probabilidad de ser reconocidos como una de estas unidades. Las dos técnicas más usadas en el proceso de clasificación son: redes neuronales y cadenas ocultas de Markov [Cole et al., 99].

Para desarrollar un sistema aplicable a situaciones reales, es necesario agregar un módulo de procesamiento del lenguaje natural que se encargue de las restricciones sintácticas, semánticas y prosódicas de la tarea.

1.1.5 Meta de un sistema de reconocimiento automático de voz

La meta principal en el desarrollo de sistemas de Reconocimiento Automático de Voz es la creación de sistemas capaces de reconocer voz de manera continua, espontánea y sin restricciones. Este objetivo aún no se ha logrado debido a las dificultades que surgen por la variabilidad en las características en la señal acústica, como se mencionó anteriormente, la cual degrada el desempeño de los sistemas de reconocimiento. Algunos de los parámetros que restringen a éstos sistemas se mostraron en la tabla 1.1.

Como cualquier tecnología innovadora, es importante que ésta se destine a aplicaciones útiles, de manera que las personas que hagan uso de dichas aplicaciones obtengan un claro beneficio al hacerlo. Deben, por lo tanto, cuidarse al máximo los detalles que hagan cómodo y agradable el diálogo con los usuarios, dado que de este diálogo depende en gran medida la aceptación (o el rechazo) de una determinada aplicación y, por extensión, de toda la tecnología que involucra.

1.2 La Tecnología Educativa

Las aplicaciones de las computadoras a la educación pueden dividirse en las siguientes clasificaciones generales [Alvarez, 99]:

Educación Asistida por Computadora: (Computer-assisted instruction (CAI)) - Utilizan la computadora para presentar lecciones completas a los alumnos. En el mercado existen muchos ejemplos de programas o CD para enseñar algún tema en particular, en el que todo el material necesario está contenido en el programa.

Educación Administrada por Computadora (Computer-managed instruction (CMI)) - Utilizan las computadoras para organizar las tareas y los materiales y para mantener registro de los avances de los estudiantes. Los materiales de estudios no son enviados necesariamente por la computadora.

Educación con Multimedia a través de Computadora. (Computer-Based Multimedia (CBM)) - Es un importante medio, aún en desarrollo, de sofisticadas y flexibles herramientas de computadoras que tienen como objetivo integrar voz, sonido, vídeo, animaciones, interacción y otras tecnologías computacionales en sistemas integrados y fácilmente utilizables y distribuíbles.

Educación mediada por Computadoras. (Computer-mediated education (CME)) - Se refiere a las aplicaciones de las computadoras que permiten el envío de materiales de aprendizaje. Incluye el correo electrónico, grupos de noticias, foros de discusión, Internet, WWW, páginas web. Es el medio con el más grande e importante crecimiento de los últimos tiempos y en este medio están basadas muchas de las potencialidades futuras de la Educación a Distancia.

1.2.1 Ambientes de Aprendizaje

El aprendizaje continuo, según la definición de la ELLI, European Lifelong Learning Initiative, es "un proceso continuo sustentativo, que estimula y da fuerza a los individuos para adquirir todo el conocimiento, valores, habilidades y comprensión que requieran durante su vida para que los apliquen con confianza, creatividad y gusto en todos los roles, circunstancias y medio ambiente en que se desarrollen" [Ayala, 99].

Un ambiente de aprendizaje es un espacio en el cual existe un problema de aprendizaje y se intenta solucionarlo de manera colaborativa y sistemática a través de interactuar con los elementos del ambiente de aprendizaje, definir el movimiento en los 3 ejes del ambiente de aprendizaje (Clarificación del problema, Mapa Personal y Aplicación) y cumplir con las distintas etapas del ambiente de aprendizaje para la solución del problema [Porras, 99].

El proceso de aprendizaje es individualizado, cada persona tiene diferentes habilidades y capacidades. En un ambiente de aprendizaje el estudiante cuenta con libertad de actuar y de mediar los elementos con los cuales interactuará. De esta forma el aprendizaje se adecua a cada persona.

1.2.1.1 Modalidades en el aprendizaje

Las modalidades concretas de los ambientes de aprendizaje, se clasifican en [Rivera, 99]:

Tutorial. En ésta modalidad se representa un material en la pantalla de la computadora y se van haciendo preguntas sobre dicho material. Se pueden hacer evaluaciones al estudiante y se le da retroalimentación.

Ejercitación y práctica. Sirve como una labor para reforzar el aprendizaje, trata de que los usuarios adquieran una habilidad sobre algo realizando ejercicios únicamente, es decir no se propone una teoría o explicación sobre el contenido de lo que se esta haciendo.

Juegos. La finalidad de ésta modalidad es que el estudiante aprenda, practique o desarrolle alguna habilidad divirtiéndose.

Simulaciones. Emplea la computadora para representar una escena cambiante en el tiempo. Permite adquirir alguna habilidad o aprender reglas para manipular un fenómeno, mecanismos o dispositivos dinámicos y complejos.

Herramientas. Son paquetes o aplicaciones que sirven para auxiliar a las tareas educativas, su finalidad no es enseñar algo sino realizar una tarea o acción específica

1.2.1.2 Componentes en el diseño de Ambientes de Aprendizaje

Formulación de propósitos y objetivos

Se debe establecer las metas y objetivos a cumplir en el ambiente de aprendizaje. Cuál es la modalidad específica en que se ubicará y en cuanto tiempo se desarrollará.

Perfil del usuario

A que personas esta destinado, el promedio de edades, necesidades y motivaciones principales.

Selección del contenido

Que material se utilizará. El uso del contenido permite la división de las tareas, en el conjunto de ideas en que se descompone y como se encadenan estas para realizar el objetivo de una lección.

Selección de estrategias de aprendizaje

Las estrategias pueden agruparse en tres sistemas educativos:

tradicionales caracterizados por una correspondencia en los componentes tecnológicos con los empleados hace una generación.

tradicionales reformados en los cuales se han introducido mejoras o adiciones a los recursos tecnológicos.

innovadores desarrollados por la tecnología educativa y caracterizados por adoptar una organización diferente en el aula con respecto al estudiante o con la organización escolar.

Diseño de Interfaz

Diseñar el modo en el que aparecerá el material de aprendizaje en la pantalla, tanto en el aspecto espacial, es decir, al colocación de textos y gráficos, como en el aspecto temporal, es decir, el tiempo de aparición del material en pantalla.

Selección y uso de medios de aprendizaje

La selección de medios, dependerá sobre todo del material con que se cuenta para cumplir el objetivo: hardware y software para lograrlos.

1.2.2 Enseñanza del lenguaje asistida por computadora

La educación asistida por computadora (CAI) ha sido una herramienta con la que los estudiantes se apoyan en el proceso del aprendizaje. En los últimos 40 años, ha habido un incremento exponencial en el uso de las computadoras en apoyo a la educación. Por medio de las computadoras el estudiante puede tener una forma de aprendizaje más sofisticada.

La inteligencia artificial simbólica ha propuesto interesantes esquemas en ICAI, pero el uso de redes neuronales ha sido propuesto muy pocas veces. En general el conocimiento para CAI es representado explícitamente con redes neuronales las cuales rara vez han sido utilizadas en ésta área [Ayala,99].

Con los avances recientes en tecnología multimedia, el aprendizaje de idiomas asistido por computadoras (CALL) ha emergido como alternativa al tentar a los modos tradicionales de suplir o de substituir la interacción directa del estudiante-profesor, tal como el laboratorio de idiomas o el *self-study*. La integración del sonido, la interacción de la voz, los textos, el vídeo, y la animación han permitido crear ambientes interactivos a ritmos individuales de aprendizaje, los cuales prometen realzar el modelo del aula de clase. Un número creciente de los editores de libros de textos ahora ofrecen software educativo de una cierta clase, y los educadores pueden elegir entre una variedad de diversos productos. Todavía, el impacto práctico del CALL en el campo de la educación de un segundo idioma ha sido algo modesto. Muchos

educadores son renuentes a abrazar una tecnología que todavía busque la aceptación de la comunidad de la enseñanza de idiomas en su totalidad [Kenning & Kenning, 90].

1.2.3 Análisis de sistemas para la enseñanza de un segundo idioma.

Diferentes grupos han desarrollado aplicaciones interesantes y útiles para el apoyo del aprendizaje de un segundo idioma.

El CSLU del OGI y el CSLU de la Universidad de Colorado han estado colaborando con educadores de *Tucker Maxon Oral School* en un esfuerzo conjunto enfocado a el entrenamiento de voz con niños con problemas de sordera. Ellos han desarrollado un Toolkit que incorpora reconocimiento de voz y facilidades de producción, así como un agente animado conversacional llamado *Baldi*. El agente es representado por un rostro en tercera dimensión que produce un lenguaje visual: movimiento de labios, lengua, ojos, cejas, durante la producción del habla mediante el uso de un sintetizador de voz. El niño puede jugar con la interfaz del lenguaje, cada lección presenta diferentes preguntas y si da la respuesta correcta entonces puede continuar.

Otro ejemplo es un software llamado *Pronunciation Power* (ver figura 1.2) el cual proporciona al usuario una serie de herramientas que le permiten aprender la pronunciación del idioma Inglés. Crea una gráfica de onda sonora de el sonido que se desea verificar, para esto, el usuario debe grabar su voz y comparar la gráfica generada con la gráfica de la "forma correcta" de pronunciación.

La desventaja de este método es que cuando hay un error en la pronunciación este no es establecido explícitamente. Se requiere de práctica, pues se debe ver la representación gráfica de la señal de voz y la de "lo correcto" y determinar como y porque ocurrió el error. Esto puede resultar relativamente fácil para una persona que tenga experiencia, pero puede resultar tedioso. Además, no se sabe en base a que criterio es elegida la representación gráfica como correcta. Una pronunciación correcta puede variar mucho en su apariencia debido a las diferencias naturales en el tracto vocal humano que puede hacer que el usuario crea que algo esta mal.

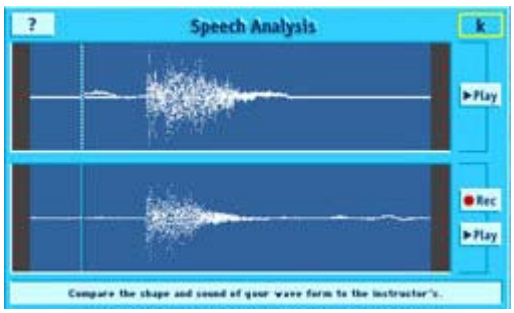

Figura 1.2 Pronunciation Power(TM). Comparación de los sonidos mediante

la representación gráfica de la señal de voz.

Otro ejemplo interesante es de la compañía *Language Connect*, la cual usa IBM *Via Voice*. El software, diseñado para la enseñanza de Inglés, ”oye” cada palabra, frase u oraciones complejas, entonces responde al usuario y da un puntaje sobre la pronunciación realizada. Este software, que utiliza reconocimiento de voz, es muy poderoso y tiene muy buen nivel de reconocimiento. Sin embargo, el puntaje que recibe el estudiante no es muy explícito, es decir, se sabe que existe un error pero no se indica en qué o dónde, y entonces no se sabe hay que corregir la pronunciación.

Otra compañía en el mercado es *Syracuse Language Systems* que en conjunto con *Dragon Systems* ha desarrollado sistemas para el aprendizaje de un segundo idioma como Inglés, Francés, Japonés, Alemán, Hebreo, Italiano y Español. Entre el software enfocado a la enseñanza del idioma español de esa compañía *estan All-in-One, Language Fun, Kids! Spanish, Let's Talk Spanish, Self-Study Spanish, Smart Start Spanish, Success in Spanish* y *Spanish Your Way*. Estos sistemas, al igual el anterior sistema, tienen Reconocimiento de Voz pero en cuanto a la verificación de pronunciación no se sabe en base a que llevan a cabo la evaluación y el puntaje, así mismo no se sabe cuales son los errores que se deben corregir en la pronunciación para subir la puntuación obtenida. En la figura 1.3 se muestra el software *Let's Talk English (TM)* que es un software para el aprendizaje del idioma Inglés.

Figura 1.3 Let's Talk English.

Learn Spanish Fluently! Es otro sistema de aprendizaje del español

con Reconocimiento de Voz y simula la comunicación interactiva. Es de la compañía *Digital River*.

Berlitz Think & Talk Spanish analizan el acento mediante reconocimiento de voz. Se escucha la pronunciación nativa de varios locutores en diferentes situaciones en las cuales se puede participar, así también se hacen lecturas y se presentan videos. La compañía es *Berlitz* y ha desarrollado sistemas de éste tipo para varios idiomas.

Figura 1.4 Software para enseñanza de español

A parte del software mencionado anteriormente existe una gran variedad de herramientas, programas y juegos que intentan ayudar a personas a aprender algún idioma. Dichas herramientas se adaptan a las diferentes necesidades del usuario. Algunas de éstas se muestran en figura 1.4.

Conclusiones

Hasta el momento hemos analizado los aspectos básicos de los sistemas con reconocimiento de voz y los sistemas educativos. Se hizo un análisis sobre algunos sistemas comerciales que existen para la enseñanza de un segundo idioma. En el siguiente capítulo se describe el problema a resolver y trabajos relacionados con este problema.

Aguas García, N. 1999. Verificación de Pronunciación Basada en Tecnología de Reconocimiento de Voz para un Ambiente de Aprendizaje. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.