Capítulo 4. Aspectos sobre el Reconocimiento de Voz.

El reconocimiento de voz consiste en convertir un flujo de palabras del lenguaje a texto. De acuerdo con las características y funciones de los reconocedores, estos pueden clasificarse como:

- Reconocedores de propósito específico: Son aquellos cuyo vocabulario está restringido por un dominio, es decir, un conjunto o subconjunto determinado como letras, números, vocales, entre otros.

- Reconocedores de propósito general: Son aquellos cuyo dominio es general, como un idioma en particular cuyas palabras no caen en un conjunto determinado.

4.1. Antecedentes del reconocimiento de voz

En esta sección se presenta un poco de historia sobre historia del reconocimiento de voz y su evolución.

En el año de 1870, Alexander Graham Bell tenía en mente el desarrollo de un aparato capaz de representar el habla de manera visible a las personas hipoacústicas. Aunque el resultado de dicho experimento no fue el esperado, este derivó en un aparato capaz de transmitir señales de voz a distancia, conocido actualmente como el teléfono [Kirschning., 1999] [Ahuactzin, 1999].

Posteriormente, en 1880, un científico húngaro llamado Tihaner Nemes solicitó el permiso para obtener la patente de una máquina que el mismo desarrolló, capaz de transcribir de forma automática una señal de voz. Sin embargo, su proyecto no pudo hacerse realidad debido a que el proyecto fue catalogado como poco realista [Ahuactzin, 1999] [Kirschning, 1999].

Los primeros intentos para reconocer la voz trataron de tomar espectrogramas del sonido en la década de los cincuenta. Las vocales eran reconocidos por su espectro gráfico. Eso era suficiente para reconocer un vocabulario limitado, como los diez primeros dígitos desde el cero al nueve, pronunciados por la misma persona.

Cuando una persona decía cada una de las diez palabras al micrófono del reconocedor de voz, extrayéndose parámetros de los patrones espectrales de cada palabra y siendo almacenados en una memoria electrónica, considerados como plantillas. Cada palabra tenía su propia plantilla (template) según fuera pronunciada por un usuario en particular. A ese proceso en que las plantillas eran elaboradas se le llama etapa de entrenamiento del reconocimiento de voz [Rodman, 1999].

Más adelante, los laboratorios de AT&T comenzaron a desarrollar la primera máquina para el reconocimiento de voz. El sistema debió ser entrenado para que pudiera reconocer las palabras de cada locutor individualmente, logrando una eficacia de 99 por ciento en el reconocimiento [Ahuactzin, 1999] [Kirschning, 1999].

Para la década de los años sesenta, los vocabularios se extendieron y se mejoró la extracción de parámetros y los algoritmos de clasificación de plantillas. El

descubrimiento más importante de la década fue Dymanic Time Warping (DTW), el cual significaba la normalización del tiempo: la duración de un utterance de prueba era ajustado para concordar con la longitud de una plantilla con la cual estaba siendo comparada.

En la década de los setenta, exactamente en el año 1971, comenzó la era de ARPA (Advanced Projects Research Agency (ARPA) de la Sección de Defensa Norteamericana, la cual desafió a las compañías americanas y a las universidades del país a desarrollar un sistema de reconocimiento de voz cuyo vocabulario consistiera como mínimo de 1000 palabras y que pudiera procesar voz continua con un error debajo del diez por ciento de error aún en un ambiente con poco ruido, probado por muchos usuarios. Gracias a las investigaciones y avances de las universidades y compañías mencionadas anteriormente, pudo contarse en esa década con algoritmos y técnicas de gran eficiencia como time warping, el algoritmo de retropropagación y el modelado probabilística [Rodman, 1999].

Después de ARPA, en la década de los ochenta, la investigación en el campo del reconocimiento de voz se enfocó en cuatro áreas: técnicas de modelado estadístico utilizando niveles mayores de conocimiento lingüístico, inmunidad al ruido y el costo. En cuanto al *modelado estadístico* podemos decir que Harpa marcó la pauta al utilizar una técnica de relación de patrones espectrales probabilística basada en los modelos de Markov. En los ochenta se estudiaron varias variaciones del método y se implementaron variaciones del método satisfactoriamente y hoy en día hay muchos sistemas de reconocimiento exitosos gracias a los modelos de Markov. El reconocimiento de patrones ya no era el enfoque principal, sino el modelado probabilística, en especial los Modelos Ocultos de Markov (HMM). La *inmunidad al ruido* comenzó a desarrollarse poco después de los proyectos de ARPA. Ninguno de los proyectos tomó en cuenta el factor ruido porque los experimentos se realizaron en ambientes silenciosos.

Al igual que con las computadoras personales, el *costo* de los sistemas de reconocimiento de voz se redujo a medida que comenzaron a ser viables para el consumo de un número cada vez más grande de usuarios. Los costos de los proyectos de investigación de reconocedores de voz en los años ochenta eran muy superiores a los costos de los productos comerciales de la actualidad. Ahora se pueden disfrutar las ventajas de emplear un reconocedor para fines médicos, educativos, de asistencia, etc. [Rodman, 1999].

Más adelante, considerando la década de los ochenta y los noventa, hicieron su aparición los sistemas de vocabulario amplio, aquellos que poseen más de 1000 palabras, bajando los costos y haciéndolos accesibles a cada vez más consumidores. Las aplicaciones de flujo continuo, donde no hay pausas significantes, e independientes del locutor, son hoy en día una realidad. Entre las empresas más importantes se encuentran: Philips, Lernout & Hauspie, Sensory Circuits, Dragon Systems, Speechworks, Vocalis, Dialogic, Novell, Microsoft, NEC, Siemens, Intel (apoyo / soporte técnico), entre otros.

4.2. Arquitectura de un Sistema de Reconocimiento de Voz.

Un sistema de reconocimiento de voz consta de los siguientes componentes:

- Extractor de características:

Una vez que recibe la señal, el extractor de características divide la señal en una colección de segmentos, aplicándole alguna técnica de procesamiento de señales para obtener la representación de las características acústicas más distintivas de segmento.

Con las características obtenidas, se construye un conjunto de vectores que constituyen la entrada al siguiente módulo [Ahuactzin, 1999] [Kirschning, 1999].

- Clasificador probabilístico:

Se crea un modelo probabilístico basado en redes neuronales, como sería el caso de los Modelos Ocultos de Markov (HMM). Una vez que se obtuvieron las probabilidades, es realizada una búsqueda para obtener la secuencia de segmentos con mayor probabilidad de ser reconocidos [Ahuactzin, 1999] [Kirschning, 1999]

4.3. Sistemas de Síntesis de Voz

La síntesis de voz es también llamada Text-To-Speech (TTS), Texto a Voz, convierte el texto en palabras audibles por medio de una voz artificial. Es un modo de retroalimentación que puede ser usado en sustitución o como complemento de mensajes visuales, creando en el usuario mayor concentración en sus tareas al no ser distraído por las ventanas de mensajes. Por eso, la síntesis de voz ha sido empleada en sistemas diseñados para personas con capacidades diferentes, facilitándoles su uso [Ahuactzin, Larios, 1999] [Kirschning, 1999].

El CSLU Toolkit posee diferentes versiones de sintetizadores en distintos idiomas aparte del inglés, como el español, italiano o el portugués. Teniendo las palabras en forma de texto, el sintetizador las procesa y ofrece su versión en lenguaje hablado.

4.4. Tipos de Síntesis de Voz.

- Concatenativa.

  Se basa en ejemplos del lenguaje humano:

  - Fonemas.
  - Palabras,
  - Unidades de longitud variable.
- Síntesis de Formantes.
- Simula de manera electrónica el habla humana usando reglas fonológicas [Kirschning, 2000].

4.4.1. Estructura de un Sintetizador de Voz (Text-to-Speech)

Los módulos de composición de un sintetizador se muestran en la siguiente figura (4.1):

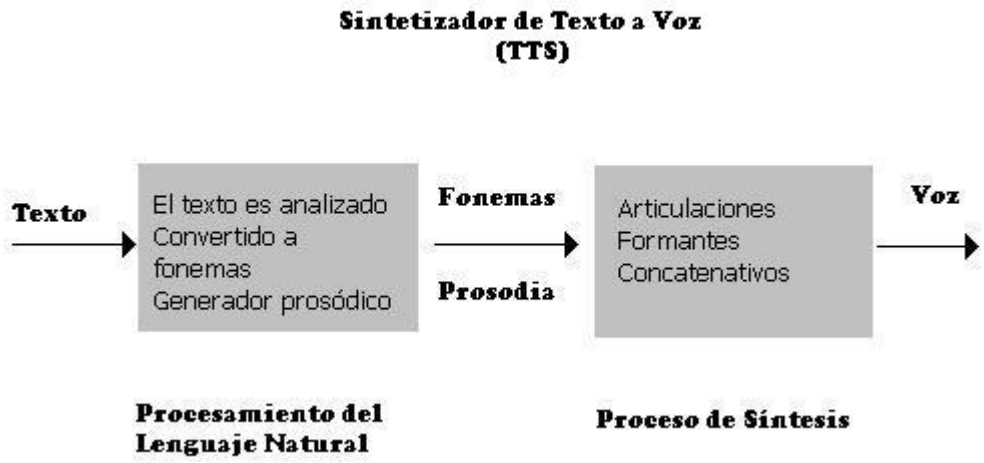

Fig. 4.1. Composición de un sintetizador de voz [Ahuactzin, 1999].

4.4.2. Módulo de Procesamiento de Lenguaje Natural

Este módulo recibe como entrada un texto que puede estar formado por palabras o frases que se interpretarán como tales para su comprensión. Posteriormente se realiza la trascripción fonética del texto que se leyó, junto con su entonación y ritmo para producir la salida de voz.

El proceso es el siguiente: Primero pasa por el Analizador de Texto, el cual toma la entrada y la transforma en una secuencia de palabras en la que ha sido eliminado el ruido, insertándole las pausas correspondientes entre frases.

Una vez que el texto ha sido analizado por el convertidor de texto a fonemas, es convertido a fonemas. Cada palabra es asociada con su trascripción fonética, es decir, con los fonemas que la componen, como por ejemplo, la palabra “texto” corresponde a /t/e/ks/t/o. Esto es necesario para que sea posible asignar la pronunciación de cada palabra de manera correcta y dar como salida una señal de voz [Ahuactzin, 1999].

Por último, el generador prosódico asigna la duración y la correcta entonación a cada fonema.

4.4.3. Módulo del Proceso de Síntesis

En éste se realiza la transformación de la información simbólica proveniente del procesamiento de lenguaje natural en la salida de la voz.

Como entrada se recibe un conjunto de símbolos debiendo ser interpretados en forma de palabras aunque no siempre tengan una asociación directa con las palabras que representan, por ejemplo, las abreviaturas, Sr., Sra., etc., o aquellas que representan en diferentes ocasiones elementos lingüísticos aparte, como los signos de puntuación que pueden influir en la pronunciación de las palabras [Ahuactzin, 1999].

4.4.4. Tipos de Síntesis

4.4.4.1. Síntesis de articulación: Consiste en emplear parámetros y factores para producir la voz como la tráquea, la posición de la lengua, tamaño de la cavidad oral, tomando en cuenta las descripciones específicas y detalladas de mecanismos fisiológicos de producción de voz y generación de sonidos en el aparato vocal [Ahuactzin, 1999].

4.4.4.2. Síntesis paramétrica: Es la formación de la voz por medio de la variación de parámetros que aplican señales armónicas, logrando que se generen sonidos semejantes a los del habla. Se incluyen filtros para generar la resonancia del tracto vocal [Ahuactzin, 1999] [Kirschning, 1999]

4.4.4.3. Síntesis concatenativa: Es aquella que genera la voz sintetizada usando la concatenación de fonemas, sílabas y palabras que fueron pronunciadas previamente por una o varias personas y después fueron guardadas en alguna base de datos [Ahuactzin Larios, 1999].

4.5. Tecnologías de Voz Existentes.

4.5.1. ¿Qué son tecnologías de voz?

Son tecnologías mediante las cuales se desarrolla software que utiliza Interfaces de Voz de Usuario (VUI – Voice User Interfaces) en vez de Interfaces Gráficas de Usuario (GUI – Graphical User Interfaces), cuyo principal dispositivo de entrada y salida es un teléfono o un teléfono celular [Correa, 1996].

Las tecnologías de voz se dividen en dos categorías:

- Texto a Diálogo (TTS – Text to Speech): Tecnología por la cual las computadoras son capaces de convertir cualquier texto en palabras audibles, es decir, sonido sintetizado. Se aplica principalmente en la telefonía, mensajería, portales de información, libros o enciclopedias.

- Automatic Speech Recognition (ASR): Tecnología usada para identificar e interpretar una palabra o una frase completa pronunciada por un usuario. Provee una forma de captura para la pronunciación del usuario, comparar las pronunciaciones aceptadas y devolver un resultado preciso. Entre sus aplicaciones están la telefonía, dictado, comandos y control, ciencias médicas y discapacidad [Correa, 1996].

4.5.2. Especificación del “Voice Browser Working Group – W3C”

4.5.2.1. VoiceXML

VoiceXML(the Voice Extensible Markup Language) es un estándar diseñado para crear diálogos de audio para ejecutar síntesis de voz, audio digitalizado, reconocimiento de lenguaje hablado, grabación de entradas de voz y telefonía. Su meta principal es emplear las ventajas del desarrollo Web y la entrega de contenidos a aplicaciones interactivas de voz.

Sus orígenes se remontan al año 1995, como un lenguaje de diseño de diálogo basado en XML para simplicar el proceso de desarrollo de aplicaciones de reconocimiento de voz para un proyecto de AT&T llamado PML (Phone Markup Language). Con la reestructuración de AT&T, Lucent y Motorola continuaron trabajando en sus propios lenguajes basados en PML [Correa, 1996].

4.5.2.1.1. Modelo Arquitectural

El modelo arquitectural asumido por la especificación tiene los siguientes componentes (ver figura 4.2):

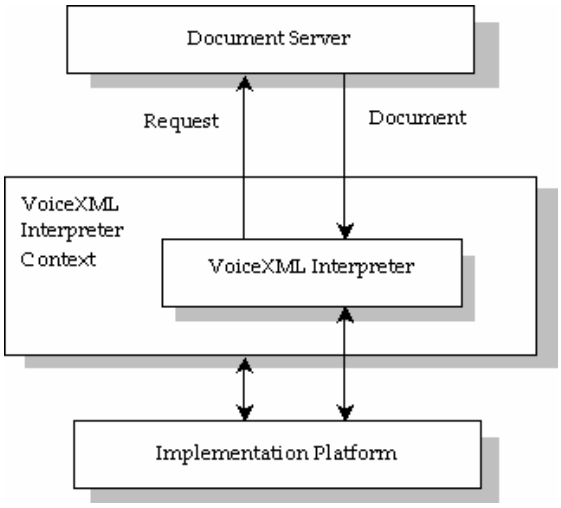

Figura 4.2. : Modelo Arquitectural [Correa, 1996].

Un servidor de documentos, como un servidor Web, procesa la solicitud de una aplicación cliente, El intérprete de VoiceXML, a través del contexto de VoiceXML intérprete. El servidor manda como respuesta documentos VoiceXML, los cuales son procesados por el intérprete de VoiceXML. El contexto del intérprete de VoiceXML puede monitorear las entradas del usuario en paralelo con el intérprete VOICEXML [Correa, 1996].

4.5.2.1.2. Metas y Alcances de VoiceXML

El alcance principal de VoiceXML es resaltar el poder del desarrollo Web y la entrega de contenidos a aplicaciones con respuesta de voz y para librar a los desarrolladores de tales aplicaciones de la programación a bajo nivel y el manejo de recursos. Permite la integración de servicios de voz con servicios de datos usando el paradigma cliente-servidor el cual es muy familiar a los desarrolladores Web. Un servicio de voz es visto como una secuencia de diálogos de interacción entre un usuario y una plataforma de implementación.

El lenguaje describe la interacción humano-computadora provisto por sistemas de respuesta de voz, los cuales incluyen:

- Salida de síntesis de voz (texto-a-voz).
- Salida de archivos de audio.
- Reconocimiento de entrada hablada.
- Reconocimiento de DTMF entrada.
- Grabación de entrada hablada.
- Control de flujo de diálogo.
- Telefonía como la transferencia de llamadas y la desconexión

4.5.2.2. Speech Recognition Grammar Specification (SRGS)

Esta especificación define la sintaxis para la representación de las gramáticas. Una gramática define el conjunto de expresiones que un usuario puede emitir o ingresar al interactuar con aplicaciones de voz [Correa, 1996].

Se pueden definir en dos formatos:

- Augmented BNF (ABNF) (Véase Fig. 4.3).
- XML

Son gramáticas libres de contexto.

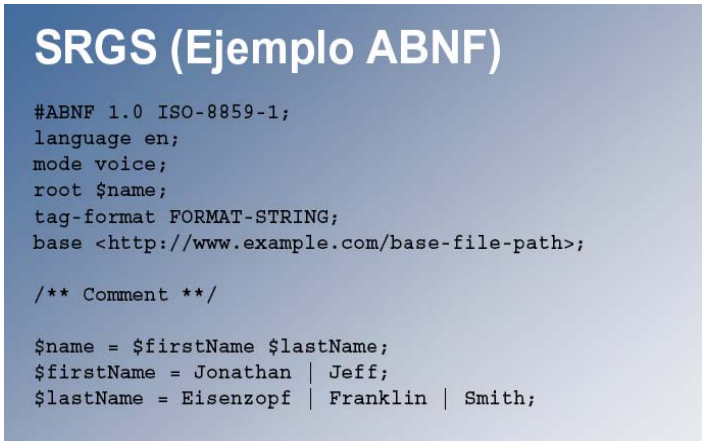

Fig. 4.3. Ejemplo de ABNF [Correa, 1996].

4.5.2.3. Semantic Interpretation Language

Describe la sintaxis y la semántica para etiquetas de interpretación durante el proceso de reconocimiento (DTMF o Speech Recognition). Fue diseñado para usarse con gramáticas SRGS.
Las notaciones son expresadas usando un subconjunto de ECMAScript y puede definirse en dos formatos: Augmented BNF (ABNF) y XML. El resultado puede ser un objeto ECMAScript o XML Speech Syntesis Markup Language

Describe la sintaxis XML usada para ayudar en la reproducción de sonido sintetizado (TTS).
El creador del sonido sintetizado puede controlar las características de voz y parámetros como el nombre, la edad, el género, el volumen, la velocidad, el énfasis y la escala [Correa, 1996].

4.5.2.4. Call Control XML (CCXML)

Describe la sintaxis XML para controlar las llamadas telefónicas para VoiceXML. Éste último y CCXML son mutuamente dependientes (Véase Fig. 4.4.). Es el complemento para VoiceXML, porque provee mecanismos avanzados como:

- Conferencia
- Habilidad para recibir eventos y mensajes externos.
- Control y administración sofisticada de múltiples llamadas [Correa, 1996].

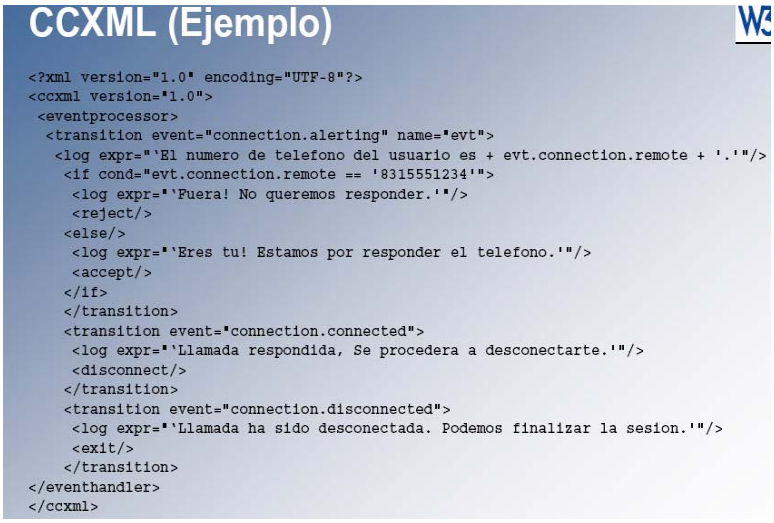

Fig. 4.4. Ejemplo de CCXML [Correa, 1996]-

4.5.2.5. Speech Applicacion Language Tags (SALT)

El standard SALT está siendo diseñado para extender lenguajes de hipertexto como HTML,XHTML y XML. El acceso multimodal permitirá a los usuarios a interactuar con una aplicación en una amplia variedad de formas: permitiéndole introducir datos usando la voz, un teclado, un ratón, y producir datos como voz sintetizada, audio, texto, video y gráficos. Cada uno de estos modos podrá ser usado independientemente o concurrentemente. Por ejemplo, un usuario puede dar un click sobre la información de un vuelo en un icono en un dispositivo y dirá “Muéstrame los vuelos desde San Francisco a Boston después de las 7 de la noche en sábado” y el navegador desplegará una página web con los vuelos correspondientes [Correa, 1996].

SALT es un conjunto más pequeño de elementos XML que mejoran los lenguajes markup existentes con una interfaz de voz. SALT puede usarse efectivamente con las características de HTML. SALT no define un nuevo modelo de programación; reusa el modelo de ejecución existente de Web para que el mismo código de aplicación pueda ser compartido a través de ciertas modalidades. Y como SALT no altera el comportamiento de los lenguajes markup con los cuales es usado, SALT tiene garantía del futuro, podrá ser usado con cualquier estándar futuro XML [Correa, 1996].

Los principales elementos de SALT son:

- Etiqueta <prompt> para configurar al sintetizador de voz y ejecutar los mensajes de salida.

- Etiqueta <reco> para configurar el reconocedor de voz a que ejecute reconocimiento y manejo de eventos de reconocimiento.
- Etiqueta <grammar> para especificar entradas de recursos gramáticales.
- Etiqueta <bind> para procesar resultados de reconocimiento en la página.

Estos elementos son activados ya sea por declaración o por programación bajo un script que corre en el lado del cliente. SALT también provee servicios telefónicos para navegadores de telefonía que ejecutan aplicaciones de solo voz [Correa, 1996].

4.6.CSLU Toolkit

Es un sistema desarrollado por Center for Spoken Language Understanding, compuesto por un conjunto de herramientas y tecnologías como el reconocimiento y la síntesis de voz muy útiles para la investigación y la creación de aplicaciones de la ciencia del habla y del lenguaje [Ahuactzin, 1999].

Incluye interfaces, dispositivos de audio, un grupo de librerías para el reconocimiento de voz, un generador rápido de prototipos (CSLUrp) y un shell de programación (CSLUsh).

4.6.1. Generador de Prototipos – RAD (Rapid Application Developer)

Es una aplicación de CSLU Toolkit en cuyo ambiente gráfico se puede desarrollar y probar un sistema de diálogo de manera rápida y sencilla, bajo un ambiente gráfico amigable. Un prototipo se crea y se diseña primero utilizando los objetos con diálogo de una barra situado al lado izquierdo del área de trabajo, pudiendo arrastrarlos a dicha área para crear la aplicación, unidos por flechas que mantienen la secuencia de ejecución, habiéndose definido previamente en cada estado el vocabulario y su gramática que el reconocedor de voz aceptará para la ejecución de determinadas tareas. Posteriormente, se compila la aplicación usando los comandos del menú o de acceso rápido para ver su ejecución y los correspondientes resultados. Puede observarse una representación de un muñeco animado que habla articulando las palabras con el movimiento de sus labios y con su expresión facial [Ahuactzin, 1999].

El ambiente fue desarrollado por el Oregon Graduate Institute of Science and Technology (OGI) y posee las herramientas que se requieren para el desarrollo de un prototipo, pues cuenta con un reconocedor de voz y un sintetizador que hacen posible una interfaz para voz [Ahuactzin, 1999].

4.6.2. Ambiente de programación CSLUsh

Se trata de un ambiente de programación desarrollado en los lenguajes C y Tcl que posee una colección de librerías de software, algoritmos y un conjunto de API's (Advanced Programming Interfaces) necesarios para la programación en el Toolkit, para el desarrollo y entrenamiento de los reconocedores. Gracias a este ambiente de programación, es posible realizar funciones complejas para el manejo de voz, lenguaje, transferencia y almacenamiento de datos, así como las aplicaciones cliente/servidor. [Aguas, 1999] [Ahuactzin, 1999].

4.6.3. Reconocedor de Voz

Toolkit cuenta con un reconocedor de voz integrado para el reconocimiento del lenguaje español mexicano para adultos, desarrollado por el grupo de investigación Tlatoa de la Universidad de las Américas, Puebla. El reconocedor de voz por defecto es aquel construido para el idioma inglés americano.

4.6.4. SpeechView

Es una herramienta de análisis de señales de voz que las despliega en forma de espectrograma, pudiéndose observar su comportamiento. Entre otras funciones de esta herramienta se encuentran la grabación de señales de voz a archivos de sonido de formato .wav y la manipulación de la voz.

4.7. Aplicaciones de Voz existentes en Domótica.

Las interfaces de voz para el control en Domótica todavía están poco desarrolladas. Las aplicaciones son todavía relativamente inmaduros y son relativamente difíciles de implementar en el contexto del hogar excepto para situaciones muy concretas como el sitio de teletrabajo o equivalente. Muchos aparatos, como teléfonos móviles, sin embargo empiezan a tener mejor y mejor reconocimiento de voz y quizás puedan ser utilizados como interface de voz para aplicaciones más amplias en el futuro [Junestrand, 2003].

4.7.1. Ventajas de una interfaz de voz

Las ventajas de tener una interfaz de voz son amplias. Permite accesar al sistema sin necesidad de elementos gráficas, lo que añade calidad y rapidez del acceso telefónico al sistema al mencionar los menús numéricos.

Este proyecto hace posible el uso de lenguaje natural, dándole libertad y confianza al usuario, usando una interfaz sencilla, sin necesidad de memorizar largos comandos particulares o una sintaxis determinada [Junestrand, 2003].

4.8. Aplicaciones de Voz para Domótica.

Existen en el mercado tanto europeo como americano diversos productos que han sido incorporados en Domótica para el reconocimiento de comandos de voz para el control y automatización de una vivienda. Algunos proyectos son de tipo académico y otros de tipo comercial.

4.8.1. HAL2000

Es una herramienta de software que facilita una integración electrónica de la casa (ver figura 4.5). Integra una gran variedad de dispositivos eléctricos y electrónicos del hogar, permitiéndole al usuario programarlos, fijar su funcionamiento y dejar que ellos interactúen con reciprocidad entre sí. Como un ejemplo, se puede programar HAL2000 para simular la presencia de alguien dentro de la casa y alejar a los extraños de tal

manera que se enciendan y apaguen las luces a voluntad, radios o cualquier otro aparato eléctrico. El sistema hace que los dispositivos interactúen entre sí, teniendo un sensor inalámbrico de detección de movimiento y hace que la PC ofrezca una respuesta a ese movimiento detectado. Puede enviar un mensaje a un teléfono móvil u otro dispositivo, encender el televisor en un canal dado, hacer sonar una alarma o sirena, encender todas las luces del jardín o de otra zona [ATV Informática-Domótica, 2005].

El sistema también ofrece otros servicios de telefonía altamente integrados. Puede recibir voz, fax, mensajes electrónicos y notificarlos en el teléfono móvil, además de transmitir información desde la red Internet, el estado del tiempo, las noticias y otra programación de televisión y mandar todo esto a un dispositivo móvil [ATV Informática-Domótica, 2005].

El sistema está adaptado para operar con la voz humana. Utiliza la tecnología de reconocimiento de voz Lernout & Hauspie. Permite usar micrófonos al aire libre, teléfono inalámbrico o llamar desde cualquier parte del mundo para operar los dispositivos de la vivienda. Utiliza frases en lenguaje natural.

Fig. 4.5. Ejemplo de operación del software HAL2000 [ATV Informática-Domótica, 2005].

4.8.2. Sicare Light

Sicare Light es un pequeño y completo mando de control con reconocimiento de voz para personas con discapacidad física [PROINSSA S.L.L, 2006].

Después de una fase de entrenamiento simple, Sicare Light interpretará las instrucciones que la persona le indique usando simplemente la voz, permitiendo el control de los aparatos que estén provistos de receptor por infrarrojos como televisores, equipos de música, sistemas de aviso de enfermería, luces, puertas, ventanas, ventiladores, calefacción, electrodomésticos, etc.

Ha sido instalado junto con el resto de accesorios domóticos, en el Hospital Nacional de Parapléjicos en Toledo, España, en dos habitaciones para asistir a cuatro pacientes.

Reconoce la voz en varios idiomas: inglés, alemán, francés, holandés, italiano, portugués y castellano. Lleva preconfigurados los códigos de infrarrojos de la mayoría de equipos existentes en el mercado, y almacena hasta 90 ordenes para controlar 23 aparatos distintos. Se pueden incorporar dispositivos mediante programación (consultar). El usuario debe realizar una fase de entrenamiento corta, repitiendo las

instrucciones que utilizará. El mando identifica las órdenes recibidas en pantalla y mediante un mensaje sonoro [PROINSSA S.L.L, 2006].

4.9. Proyectos de Investigación para desarrollo de Interfaces de Voz domóticas.

En la Universidad de Sevilla, en España, un grupo de investigación creó una interfaz de voz en lenguaje natural para un entorno domótico. Tanto la investigación como el desarrollo del sistema fueron financiadas por el proyecto Dhomme, uno de los proyectos del V Programa Marco Europeo. [Casadomo, Soluciones, 2005].

La arquitectura utilizada para este proyecto fue la arquitectura de agentes inteligentes distribuidos. Cada uno de los dispositivos controlados es un agente que contribuye al sistema global mediante su funcionalidad básica (lo que se convierte en las funciones primitivas del entorno). Por decir, una lámpara puede ser encendida o apagada. Estos agentes están conectados usando la propia red eléctrica de la casa, por lo que no es necesario instalar un cableado nuevo y específico. Es compatible con los estándares X10 y Lonworks [Casadomo, Soluciones, 2005].

El núcleo del sistema se encuentra en una computadora conectada a la red eléctrica con conexión X10 o Lonworks. En la computadora están los agentes que se encargan de la manipulación del lenguaje natural, como el reconocimiento de voz, comprensión del lenguaje natural, generación del lenguaje y síntesis del habla, así como agentes específicos de la administración del conocimiento, permitiendo la configuración, gestión y monitorización del entorno domótico y manejo de acciones que controlan los dispositivos específicos y añaden funcionalidad extra a las capacidades básicas del sistema, por decir, se encarga de modular el volumen de la televisión o del radio cuando suena el teléfono [Casadomo, Soluciones, 2005].

El equipo también desarrolló una casa virtual, donde es posible observar el rendimiento del sistema, auxiliado por una maqueta de demostración con dispositivos reales que responden a los comandos del usuario.

La implementación permite el encendido y apagado de todos los dispositivos conectados al sistema, además pregunta sobre su estado actual, su cantidad o existencia en las habitaciones o grupos de habitaciones (ver fig. 4.6).

El sistema puede controlar cualquiera de los dispositivos del entorno domótico. La implementación actual permite encender y apagar todos los dispositivos conectados al sistema, además de preguntar sobre su estado actual, su cantidad o su existencia en las distintas habitaciones o grupos de habitaciones:

Los dispositivos cuentan o pueden contar con descriptores que los identifiquen. El sistema puede además informar al usuario de posibles errores conceptuales.

Fig. 4.6. Prototipo virtual del proyecto de la Universidad de Sevilla, España [Casadomo, Soluciones, 2005]

4.10. Aplicación de CSLU Toolkit al proyecto de investigación

En pruebas anteriores Toolkit ha demostrado que es capaz de soportar un amplio vocabulario y reconocerlo correctamente la mayoría de las veces y puede interactuar óptimamente con otros lenguajes de programación de alto nivel como Java y C++, tanto en plataforma Windows como en Sun o en UNIX. Por tal razón se escogió trabajar con él. El proyecto presenta un vocabulario corto de prueba con términos relacionados a la domótica de acuerdo a la simulación que podrán interactuar con el usuario a través de comandos breves y específicos pronunciados por él en el módulo de comandos domóticos. Por otro lado, la síntesis de voz proveerá las respuestas auditivas de las interacciones con la interfaz en Java para el módulo de monitoreo. Aunque se pretende que los módulos mencionados anteriormente trabajen de manera simultánea, esto en la práctica no será posible debido a que CSLU RAD Toolkit tiene ciertas limitaciones como el no correr más de un módulo en paralelo y la interacción entre Java, C++ y RAD también es limitada.

4.11. Conclusiones

En este capítulo se presentan las tecnologías de voz más importantes y de mayor uso internacionalmente. Se puede observar que algunas apoyan principalmente al ambiente web o a la telefonía y otras son de propósito general, como lo es CSLU Toolkit, de Colorado, el cual se usará ampliamente en este proyecto y el cual brinda tanto reconocimiento como síntesis de voz, así como el marcado a un teléfono celular o a un teléfono fijo. El nivel de reconocimiento que proporciona el Toolkit es muy bueno, salvo en algunas ocasiones en que no reconoce bien el vocabulario porque se necesita calibrar el micrófono a la voz de un usuario determinado, mayormente cuando el vocabulario es de propósito general.

Asimismo se presentan los sistemas y proyectos más importantes que hay internacionalmente con respecto al procesamiento de voz en Domótica, algunos muestran buenos resultados y están en vías de ser mejorados y por otra parte, muestra que aún falta investigación en nuestro país al respecto, aunque el panorama luce esperanzador.