UNIVERSIDAD EUROPEA DE MADRID

ESCUELA SUPERIOR POLITÉCNICA

MÁSTER OFICIAL EN HOGAR DIGITAL, INFRAESTRUCTURAS Y SERVICIOS

PROYECTO FIN DE MÁSTER

Estudio de la integración de las tecnologías de reconocimiento de voz para el control y gestión del Hogar Digital.

Fernando Martín de Pablos

	**MASTER EN HOGAR DIGITAL, INFRAESTRUCTURAS Y SERVICIOS.**	**PROYECTO FIN DE MASTER**	Universidad Europea de Madrid Laureate International Universities
	Fernando Martín de Pablos	Estudio de la integración de las tecnología de reconocimiento de voz para el control y gestión del Hogar Digital.	

TÍTULO DEL PROYECTO:

Estudio de la integración de las tecnologías de reconocimiento de voz para el control y gestión del Hogar Digital.

NOMBRE ALUMNO: Fernando Martín de Pablos.

NOMBRE DIRECTOR: Miguel Roser Ballester.

FECHA DE PRESENTACIÓN: 23 de septiembre de 2.008

CALIFICACIÓN OBTENIDA:

RESUMEN DEL PROYECTO:

Estudio y análisis de técnicas y aplicaciones actuales en el reconocimiento de voz. Valoración de dichas técnicas para su implementación con los sistemas comerciales actuales de Hogar Digital. Diseño del interfaz multimodal incorporando el reconocimiento y la síntesis de voz.

Formulación de soluciones técnicas para la instalación e integración (tipología y ubicación de dispositivos) de los sistemas de voz indicando ventajas y desventajas de cada solución.

Análisis de sistemas comerciales de control por voz indicando sus puntos fuertes y sus inconvenientes. Estimación económica genérica de las soluciones presentadas. Propuesta de diseño de sistema de control por voz según el estudio realizado y los objetivos planteados inicialmente.

ABSTRACT:

Study and analysis techniques and applications in speech recognition. Valuation of such techniques for its implementation with commercial current systems of Digital Home. Multimodal interface design incorporating the recognition and voice synthesis.

Formulation of technical solutions for the installation and integration (type and location of devices) of speech recognition and synthesis systems indicating advantages and disadvantages of each solution.

Analysis of the automatic speech recognition and synthesis commercial systems indicating their strengths and weaknesses. General economic estimates of the solutions presented. Proposed design automatic speech system under study and set objectives initially.

PALABRAS CLAVE:

Control por voz. Hogar Digital. Reconocimiento de habla natural. Síntesis de voz. Modelos ocultos de Markov. Fagor, Proinssa, Personica, Easy Life, Indisys.

GLOSARIO DE TÉRMINOS.

En el documento se incluyen hipervínculos que enlazan con sitios Web donde se explican los términos que aparecen como acrónimos. No obstante, se incluyen algunos de ellos en este apartado:

AD: Analógico-Digital (conversión de señales).

*ASR: (Automatic Speech Recognition).*Reconocimiento automático del habla.

BIT RATE: Bits por segundo (tasa binaria de transmisión).

DA: Digital- Analógico (conversión de señales).

DTW: (Dynamic Time Warp). Alineamiento dinámico del tiempo.

GPS: (Global Positioning System): Sistema de posicionamiento global.

HTK: (Hidden Markov Model ToolKit). Herramienta para modelado de modelos ocultos de Markov

Hz: hercio (ciclo por segundo)

ICT: Infraestructura Común de Telecomunicaciones.

IPA: (International Phonetic Alphabet). Alfabeto fonético Internacional.

LAN: (Local Area Network). Area de red local.

*PCM: (Pulse Code Modulation).*Modulación por código de pulsos.

PDA: (Personal Digital Assistant). Asistente Digital Personal.

PLC: (Power Line Comumunications). Comunicación por líneas eléctricas.

SNR: (Signal to Noise Ratio). Relación Señal Ruido.

SAMPA: (Speech Assessment Methods Phonetic Alphabet).Método de catalogación del habla por medio del alfabeto fonético.

UPNP (Universal Plug aNd Play). Sistema universal de conexionado y funcionamiento.

WLAN: (Wireless Local Area Network). Area de red local inalámbrica.

AGRADECIMIENTOS:

Estudiar y trabajar es duro. Durante los dos años que ha durado el máster, tener los jueves, viernes y sábados ocupados implica dejar de hacer algunas cosas. Doy las gracias a Lola por haber sido sufridora pasiva en este tiempo y haber sido capaz de darme ánimos para llegar hasta el final.

Gracias también a mi tutor de proyecto, Miguel Roser por orientar y dirigir diligentemente este proyecto fin de máster.

Septiembre de 2.008

ÍNDICE

1 RESUMEN EJECUTIVO

En este proyecto se aborda el estudio del control por voz del Hogar Digital teniendo en cuenta sus diferentes tipos de utilización en función del sistema utilizado y de su integración con las tecnologías de control domótico actuales. Se ha considerado en este trabajo el sistema de control por voz como un interfaz adicional a otros ya existentes en el Hogar Digital (pulsadores, pantallas táctiles, PDAs u ordenadores).

Desde hace años, el crecimiento de los sistemas de reconocimiento de voz ha ido en aumento. En un principio se comenzaron a usar voces pregrabadas con menús de opciones para aplicaciones telefónicas automáticas. Estas indicaban las opciones que el usuario podía elegir pulsando los botones del teléfono y mas tarde como reconocedores de comandos de voz, en los que usuario podía indicar cualquiera de las palabras del menú de voz. Posteriormente las aplicaciones telefónicas se han desarrollado considerablemente permitiendo una mayor flexibilidad en el reconocimiento y en la síntesis de voz. A su vez, en el mercado informático, se han comercializado programas de reconocimiento y de síntesis de voz que se ejecutan sobre ordenadores personales, tanto comerciales, como de software libre. Estos programas permiten el dictado automático de documentos, el control de los sistemas operativos y la navegación WEB mediante órdenes de voz. Son independientes del locutor y permiten el entrenamiento previo del sistema, incrementando la precisión en el reconocimiento ya que el usuario puede corregir el error cuando este se produce. De esta forma, el sistema aprende dinámicamente, y cuanto más se usa, mejor es su eficiencia. La ventaja de estas técnicas para el usuario es la rapidez en la generación de documentos y la comodidad en el manejo de los sistemas operativos. Los inconvenientes que suelen encontrarse en muchas de las aplicaciones de control por voz, son el entrenamiento previo, que implica una pérdida de tiempo inicial y los fallos en el reconocimiento, que implican un retraso en el trabajo que se está realizando para indicar al sistema cuál es la palabra correcta.

Los sistemas de reconocimiento de voz están basados, o bien en reconocimiento por comparación de patrones o bien en decodificadores acústico fonéticos. Estos últimos son muy complejos y no se usan en la práctica. De las técnicas de procesado basadas en patrones cabe destacar, por un lado, las que se apoyan en el uso de redes neuronales, y por otro, aquellas que realizan el procesado según los modelos ocultos de Markov. Existe un tercer tipo que realiza una combinación de ambas técnicas. En cualquier caso necesitan un entrenamiento previo del sistema para que el corpus de voz (base de datos) vaya adquiriendo más información y mejore la tasa de acierto del reconocedor.

Una reciente aplicación en las técnicas de voz es la de reconocimiento del hablante. Esta característica detecta una huella biométrica de cada persona, el timbre, siendo útil en sistemas de comprobación de identidad o para distinguir usuarios y aplicar perfiles de uso diferentes en el control de sistemas.

La otra parte del diálogo con la máquina, el otro sentido de la conversación, es la generación de voz artificial, (text to speech) o síntesis de voz. En la actualidad, los sintetizadores de voz consiguen unas voces ” muy reales “ que simulan emociones y se

acercan cada vez más a la forma de hablar humana. Los sintetizadores de voz permiten la personalización, pudiendo elegir tanto el género, como el acento o la velocidad del habla de la voz sintética.

Un tema a tener en cuenta en la comunicación hombre-máquina es el diseño del interfaz. Este debe ser amigable y natural y se debe adaptar a la forma de actuar de las personas. Para ello se estudia, como un paso posterior al reconocimiento de voz, la gestión inteligente del diálogo, un área de la inteligencia artificial que trata hacer natural la conversación entre el hombre y la máquina, extrayendo información mediante análisis sintáctico y semántico y relacionando el contexto de la conversación. Idealmente, además, el sistema debe ser capaz de adaptarse a diversos usuarios en función de su nivel de conocimiento del sistema en sí mismo, de sus preferencias anteriores, o del contexto en el que se encuentre. También se trata de imitar algunos fenómenos típicamente humanos: confirmaciones durante la conversación, inicio de conversación menos fluido, etc. Debe tener en cuenta que los diálogos entre humanos son variables con interrupciones frecuentes, solapamientos o frases incompletas o no estructuradas correctamente. La interacción con la máquina debe ser estructurada para que los objetivos del gestor de diálogo se realicen correctamente.

El control por voz en el hogar digital puede realizarse de formas diferentes según el tipo de elementos captadores de voz a utilizar, y de la forma de transmitir las señales de audio entre los elementos que componen el sistema de control por voz. Desde el punto de vista técnico, el conexionado más simple es el realizado en estrella desde la central hasta cada elemento emisor o receptor. Esta conexión se puede realizar con señales analógicas o digitales. En este caso en el que se transmiten las señales digitalmente, se consigue aumentar la longitud del cableado respecto a señales analógicas sin perdida de calidad. El coste económico es mayor en este segundo caso ya que debe utilizar conversores A/D y D/A en el sistema de control y en los dispositivos hardware emisores respectivamente.

Una tipología recomendada es el uso de redes LAN para el transporte de las señales entre la central y los sistemas auxiliares de captación y emisión. En cada sistema de captación y emisión será necesario añadir un adaptador al BUS (tarjeta de red).

Entrando en las técnicas de captura y transmisión de la voz para aplicaciones de reconocimiento de voz aplicadas al Hogar Digital se comprueba que unos valores óptimos pueden ser:

- Ancho de banda señal de voz: 8.000 Hz.
- Filtrado mediante paso banda con frecuencias de corte de 100Hz y 8.000Hz.
- Frecuencia de muestreo: 16 KHz
- Codificación: 10 bits
- Bit rate mínimo que debe asegurar el BUS: 1,6 Mbit/s

Según estas premisas, analizando los sistemas domóticos comerciales para el control del Hogar Digital las conclusiones a las que ha llegado son que la mayoría de los sistemas

no soportarían la inclusión de una carga de datos con un Bit Rate alto en sus buses. Por tanto es conveniente que la aplicación de control por voz disponga de un bus independiente del sistema domótico.

Sólo en el caso en que el bus domótico funcione sobre una red LAN, WAN o PLC, se asegura un correcto funcionamiento. En estos casos se dispone de una velocidad binaria de BUS más que suficiente para el transporte de las señales del sistema de voz e incluso permite compartir dicho BUS con otros sistemas. En vivienda ya construida en la que resulta difícil introducir cableado nuevo, se pueden utilizar adaptadores de línea eléctrica-Ethernet (PLC) para la extensión de la red LAN.

Se ha realizado también un estudio de las ventajas e inconvenientes según el tipo de conexión: cableada o inalámbrica y de la disposición de los elementos captadores de audio: fijos o móviles y cuyas conclusiones se recogen en la tabla 2 indicando cada caso. De igual forma se valora el problema de la incorporación de cableado adicional para llevar alimentación eléctrica a los dispositivos emisores y receptores y se considera la utilización del cableado eléctrico como portador de señales mediante PLC. De esta forma se podrían instalar los elementos captadores y emisores en el techo de las estancias.

Se llega a la conclusión que el sistema de control por voz debe interconectarse con el sistema domótico mediante el uso de un interfaz o pasarela que permita el tráfico de datos en ambos sentidos.

Un problema común en el reconocimiento dentro de una vivienda es la presencia de ruido ambiente no deseable procedente de fuentes comunes en la vivienda: electrodomésticos, equipos de audio y video, mascotas, etc. Otro efecto no deseable es la reverberación de las estancias. Para reducir la influencia de estos efectos recomienda un estudio en la instalación de los elementos captadores: zonas de mayor probabilidad de reconocimiento del captador relacionadas con la ubicación del mobiliario de cada estancia y su utilidad. Con estas consideraciones se consigue mejorar sustancialmente la eficiencia del reconocedor, punto débil del sistema. En ocasiones el sistema fallará y será necesario actuar sobre la fuente de ruido: cerrar una ventana, bajar el volumen del televisor o mandar callar a la mascota. En estos casos el propio sistema debe ser capaz de realizar estas acciones si dispone de los actuadores necesarios: control del volumen de dispositivos de audio y video o motores de persianas. En otros casos debe ser capaz de indicar al usuario que hay un ruido que impide la correcta recepción y que sea el propio usuario el que trate de reducirlo.

Es muy importante dar un nombre al sistema como si se tratara de una persona. Si el sistema se humaniza, la interacción resultará más natural para los usuarios. Desde el punto de vista técnico también resulta positiva la introducción de una palabra de atención y que permite “despertar” al sistema con una palabra clave que a su vez sirve como ajuste de los sistemas captadores de audio. Es recomendable el uso como palabra de atención de nombres poco comunes y de más de dos sílabas. También se han estudiado y valorado positivamente la inclusión de aplicaciones de verificación del hablante, localización del usuario en la vivienda y el procesado de órdenes simultaneas por diferentes usuarios localizados en puntos diferentes de la vivienda.

Se ha realizado un estudio de sistemas de control por voz y en algunos casos también de síntesis de voz de diferentes marcas comerciales: FAGOR, PROINSSA, PERSONICA, INDISTSYS e EASY LIFE. Se describe la forma de funcionamiento de cada sistema y se valoran en cada caso las ventajas e inconvenientes. También se ha estimado económicamente el precio de los sistemas en función de sus características. Con los objetivos de estudio y análisis cumplidos se realiza una propuesta de diseño de un sistema de control por voz ideal basado en los puntos fuertes encontrados en los diseños analizados. Este diseño es perfectamente realizable a día de hoy respecto al estado de la técnica.

Durante el presente trabajo se han encontrado líneas de trabajo futuras relacionadas con algunos aspectos teóricos o técnicos que requieren un estudio pormenorizado para su correcta valoración. Algunos ejemplos se mencionan a continuación:

- Diseño de un sistema centralizado de control por voz, con una unidad de procesado que realice la síntesis de voz y haga de interfaz con el protocolo domótico usado.

- Optimización en la transmisión de datos de audio mediante red LAN, WAN o PLC, es decir compatibilidad TCP/IP.

- Estudio y diseño de un sistema captador de voz formado por un hardware que realiza parte del preprocesado de audio: filtrado paso banda, ajuste de ganancia, reducción de ruido y cancelación parcial de ecos. Debe contar con al menos dos entradas para dos micrófonos omnidireccionales instalados en cada estancia. El audio es empaquetado siguiendo las directrices que establece el protocolo UPnP.

- Sistema emisor formado por un hardware que amplifica y reproduce a través de un altavoz el audio que envía la unidad central en forma de streaming siguiendo igualmente las directrices del protocolo UPnP.

Para finalizar indicar que los objetivos planteados en el anteproyecto se han alcanzado satisfactoriamente.

2 INTRODUCCIÓN

Este proyecto fin de Máster trata sobre el uso de la voz para el control de dispositivos en el Hogar Digital, entendido este último, como aquel que es capaz de integrar diferentes tecnologías: domótica, telecomunicaciones, gestión energética, seguridad, accesibilidad y ocio para el beneficio de sus habitantes. En función de las personas que lo habitan las necesidades pueden ser completamente distintas. En el caso que nos ocupa, se considera el control por voz como un interfaz adicional a otros ya existentes, aceptados e integrados en el Hogar Digital (pulsadores, pantallas táctiles o PDAs u ordenadores). Se aborda el estudio del control por voz del Hogar Digital teniendo en cuenta sus diferentes posibilidades de utilización e integración.

Se trata en esta memoria de dar una visión global de las tecnologías existentes en el reconocimiento de voz con un breve repaso histórico al desarrollo de las diferentes técnicas y sistemas, describiéndolos brevemente y clasificándolos en función de sus cualidades para centrarnos posteriormente en su uso aplicado al Hogar Digital. El objetivo final del trabajo es ofrecer una idea completa de todos los procesos, agentes, tecnologías y condiciones de instalación que hay que conjugar para que la implantación del control por voz tenga éxito. Se basa mucha de la información contenida en este documento en la recogida durante los dos años de master en Hogar Digital. Otra información proviene de mi experiencia en sectores de la ingeniería de telecomunicaciones y en el diseño y dirección de obra de ICT. El resto de la información se basa en estudios de otros profesionales publicados en Internet. Para la parte de las tecnologías existentes, en el caso de INDISYS y de EASY LIFE la información proviene de las propias empresas a las que he visitado para interesarme y aprender de sus sistemas. Aprovecho para agradecer su tiempo e interés, atenderme y resolver mis dudas.

El documento dispone de hipervínculos que enlazan con sitios Web donde poder ampliar información de cada una de las palabras o expresiones que aparezcan subrayados y en color azul. Por tanto se recomienda disponer de conexión a Internet y consultar este documento de forma electrónica si se pretende ampliar información mediante la navegación por los enlaces asociados. Si el lector dispone de conocimientos previos relacionados con el procesamiento de audio, el reconocimiento del habla y el Hogar Digital puede directamente pasar al punto 2.4. Si no es así, se recomienda una lectura desde el principio del documento para afrontar adecuadamente la lectura del resto del trabajo.

La estructura del proyecto se basa en capítulos y subcapítulos. Se ha tratado de ir de lo general a lo concreto, primero dando una visión general, incluyendo una breve referencia histórica de los sistemas existentes, y posteriormente centrando el estudio en el mercado residencial. Una vez analizadas las posibilidades de integración, se pasa al estudio de los sistemas comerciales, que se valoran y estiman en puntos posteriores y se finaliza con propuestas y conclusiones en función del recorrido de investigación realizado.

Comenzamos con un ejemplo que nos sirve para ilustrar dos situaciones en que se dispone de sistemas de Hogar Digital integrados en la vivienda, que bien pudieran ser reales, e incluso pueden resultarle cercanos al lector:

El primer caso es una vivienda habitada por una familia con niños en edad escolar y con los dos padres trabajando fuera de casa. En este caso parece claro que será necesario, entre otras cosas, el control parental para la gestión de los hábitos y tareas de los niños cuando los padres no están en casa: control de las horas de estudio, acceso a internet, televisión y videojuegos.

Por otro lado, en ese mismo edificio o urbanización, sea otra vivienda, de mismas características constructivas, habitada por un matrimonio de personas jubiladas, con algún pequeño problema de movilidad debido a la edad y que realizan actividades de ocio en el hogar. Aquí puede ser conveniente disponer de automatizaciones para subir y bajar persianas y toldos además de contar con un gestor que indique, por ejemplo, las horas de las tomas de los medicamentos de cada persona.

En ambos casos la vivienda ha de contar con un sistema de gestión, o inteligencia central que sea capaz de "recordar" las órdenes que debe ejecutar:

En el caso de la familia con niños:

1. No permitir encendido de televisión hasta las 19:30 h.
2. No permitir el acceso a Internet a sitios catalogados para adultos.

En el caso del matrimonio de jubilados:

1. El señor toma varias pastillas al día. Se recuerda en cada comida que pastilla debe tomar. Además, se controlan las pastillas que tiene cada caja y se informa con antelación suficiente para reponerlas.
2. La señora ve la televisión en verano después de comer, a partir de las 16;00 h en el comedor. El sistema de control baja las persianas del comedor y conecta el aire acondicionado a la temperatura deseada.

En ambos casos el sistema de gestión ha controlado los dispositivos asociados a diferentes servicios del Hogar Digital. La "programación previa" de dichos dispositivos ha tenido que realizarse basada en una serie de condiciones que los usuarios previamente han definido. En el caso del control parental, los padres han decidido los horarios de estudio y los sitios Web que pueden visitar sus hijos. En el caso del matrimonio de jubilados han decidido la forma de actuar del hogar según sus hábitos.

El ejemplo anterior ilustra lo que es capaz de realizar un Hogar Digital programado para realizar acciones asignadas previamente en función de los hábitos de vida. Pero la realidad es que no todos los días hacemos lo mismo. En estos casos debemos contar con un Hogar Digital que se adapte a nuestra forma de vivir y no al contrario. Estos casos deben estár previstos en la mayoría de los sistemas avanzados de control del hogar para permitir que el sistema "aprenda" nuevos hábitos o casos excepcionales.

Siguiendo con el ejemplo:

Ese día la madre ha regresado antes a casa, se ha sentado con sus hijos a hacer los deberes. Los han acabado pronto y como premio, les deja jugar hasta la hora de la cena con la consola de videojuegos. Son las 19:00 h. El sistema está programado para no activar la toma de corriente de la consola hasta media hora después.

La señora jubilada recibe a sus hijos que vienen con la familia a comer un día de diario que están de vacaciones. Se prolonga la comida y hacen sobremesa en el comedor. El sistema está programado para bajar la persiana y conectar el aire acondicionado a las 14:00h.

En estos casos se ha roto la rutina. Es necesario, que o bien el sistema detecte el cambio de rutina, o bien el usuario se lo haga saber. Para ello la madre, que cuenta con privilegios de administradora en el hogar, indica al sistema que se permite jugar a los niños antes de la hora prevista. La señora jubilada igualmente hace saber del cambio de rutina y activa el aire acondicionado en el comedor antes de comer.

La interacción con el hogar, el decirle *qué* debe hacer y *cómo* de una forma sencilla es una de las partes más importantes para conseguir su implantación definitiva. Para comunicarnos con el Hogar Digital es necesario el uso de interfaces: mandos a distancia, pantallas táctiles, pulsadores control por voz, control gestual, etc. Estos interfaces de control y gestión teniendo en cuenta su usabilidad pueden ser la "killer application" del Hogar Digital.

Volviendo al ejemplo, surgen dos posibilidades, sabiendo que el Hogar Digital dispone de interfaces multimodales, es decir que la misma acción puede realizarse de formas diferentes.

Primera posibilidad:

La madre de los niños accede al sistema de control través de la pantalla táctil validándose con su usuario y contraseña, y mediante las menos pulsaciones posibles, accede al control de la activación manual del enchufe que controla la videoconsola.

La señora jubilada igualmente accede a la pantalla táctil hasta llegar al menú de escenas programadas y desactiva temporalmente la escena de "TV después de comer".

Segunda posibilidad:

Ambas viviendas disponen del interfaz de control por voz. En ese caso, la madre dice:

"hogar, activa el enchufe de la videoconsola en el cuarto de juegos de los niños".

Y señora dice:

"hogar, hoy comemos en el comedor a las tres".

En las dos posibilidades planteadas el hogar va a ejecutar las mismas acciones, activará el enchufe de la videoconsola en el caso de la madre con los niños, y encenderá el aire acondicionado antes de la hora y no bajará la persiana automáticamente, en el caso de la señora.

El segundo interfaz, el control por voz, a priori parece muchísimo más simple ya que no hemos tenido que aprender a manejarnos por menús y familiarizarnos con el interfaz de pantalla táctil. En el control por voz tanto la madre como la señora han hablado con el sistema en su propio lenguaje, facilitando el acercamiento y humanizando el control del hogar.

El sistema puede responder en ambos casos dando una confirmación de la correcta interpretación de las órdenes. Esta confirmación podrá realizarse con síntesis de voz o simplemente mostrando en la pantalla la nueva configuración. Las ventajas e inconvenientes de cada caso se comentarán posteriormente.

Estos dos ejemplos que se han llevado en paralelo pretenden ilustrar una de las tantas posibilidades del control por voz del Hogar Digital. En puntos posteriores se tratará de estudiar el control desde el punto de vista técnico y desde el punto de vista de usuario. En ambos casos siempre se intenta dar un enfoque eminentemente práctico cara a la implantación definitiva de estos sistemas en el Hogar Digital. También se describirán las iniciativas actuales existentes a nivel de investigación y desarrollo y a nivel de comercial.

El *reconocimiento automático del habla*, también llamado *reconocimiento de voz o, Automatic Speech Recognition (ASR)* es una parte de la inteligencia artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y máquinas. Estos sistemas deben conjugar información procedente de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), y en presencia de ambigüedades, incertidumbres y errores inevitables, obtener una interpretación aceptable del mensaje acústico recibido. Esta tarea de reconocimiento no es sencilla Diariamente experimentamos momentos de ambigüedades en el reconocimiento del lenguaje, una mala recepción debida a ruido en el ambiente, señales interferentes, mala pronunciación por parte del locutor o desconocimiento del contexto. Ciertas frases no captadas correctamente son en ocasiones reconstruidas por nuestro cerebro básicamente por nuestro "entrenamiento" durante edades tempranas y por nuestra experiencia social posterior durante años. Estas acciones que realizamos diariamente de forma natural suponen un reto muy importante a la hora de modelar un sistema eficiente de reconocimiento de voz.

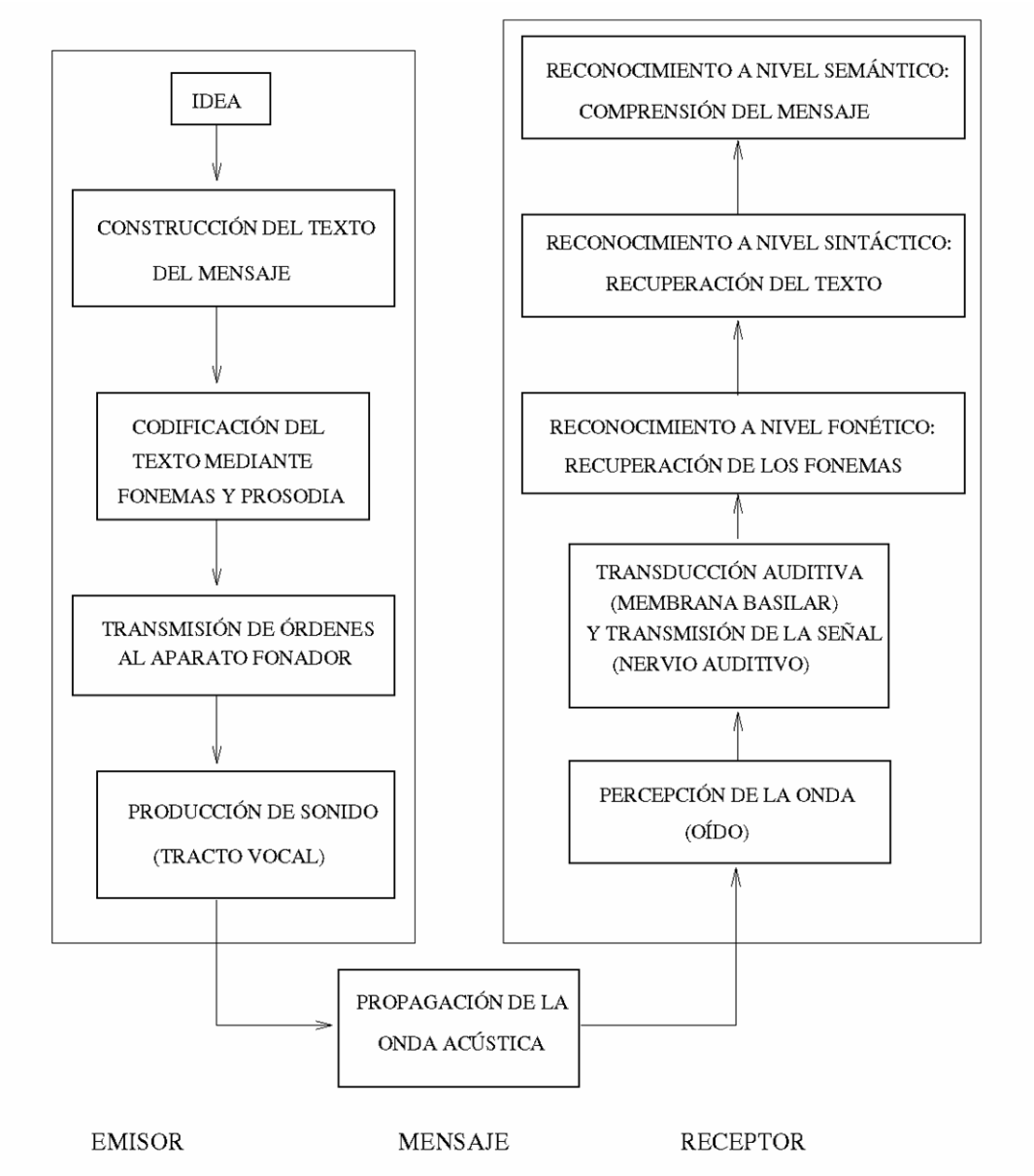


FIGURA 1

Se indican en la figura 1 los procesos implicados en la comunicación oral entre dos personas, existiendo un emisor que pretende comunicar una idea en forma de mensaje, un receptor que debe ser capaz de comprender dicho mensaje y un canal por el que se propagará la onda acústica. El canal podrá ser el aire únicamente o bien puede ser algún medio de transmisión en el que previamente se ha realizado una conversión mediante los transductores apropiados (micrófono y altavoz en una comunicación telefónica).

Los sistemas comerciales han estado disponibles desde la década de los noventa. A pesar del aparente éxito de estas tecnologías y de su integración en los ordenadores personales, la realidad es que muy pocas personas utilizan el sistema de reconocimiento del habla en su trabajo diario con su ordenador. La mayoría de los usuarios utilizan el ratón y el teclado para editar y generar documentos porque les resulta más cómodo y rápido aunque realmente podemos hablar a más velocidad de la que tecleamos. Lo que se comprueba es que con el uso combinado de varios interfaces: teclado, ratón y el reconocimiento del habla, el trabajo es mucho más efectivo.

2.1 EVOLUCIÓN HISTÓRICA

En 1870 Alexander Graham Bell quería construir un sistema que hiciera el habla visible a las personas con problemas auditivos. El resultado no fue exactamente el deseado pero sin duda revolucionó las comunicaciones desarrollando el teléfono. Unos años más tarde, en 1880's Tihamir Nemes solicita una patente para desarrollar un sistema de transcripción automática que identifica secuencias de sonidos y los imprime (texto). Fue rechazado como "proyecto no realista". Pasados 30 años AT&T Bell Laboratorios construyen la primera máquina capaz de reconocer voz basada en plantillas de los 10 dígitos del Inglés. Requería un complicado ajuste a la voz de la persona que lo iba a utilizar pero una vez logrado tenía un 99% de aciertos. A partir de entonces surge la idea de que el reconocimiento de voz es posible. La mayoría de los investigadores que estaban trabajando en el reconocimiento de voz en la década de los 60 perciben que el proceso es mucho más complicado y sutil de lo que habían anticipado. Por lo tanto empiezan a reducir los alcances y se enfocan a sistemas más específicos:

- Dependientes del locutor.
- Flujo discreto de habla (con pausas entre palabras)
- Vocabulario pequeño (menor o igual a 50 palabras)

Estos sistemas empiezan a incorporar técnicas de normalización del tiempo (minimizando la diferencia entre diferentes velocidades del habla) y ya no es necesaria una exactitud perfecta en el reconocimiento.

IBM y CMV trabajan en reconocimiento de voz continuo pero no se ven resultados aceptables hasta la década de los 70. A Principios de los 70 se produce el primer de reconocedor de voz, el **VIP100** de Threshold Technology Inc. Utilizaba un vocabulario pequeño, dependiente del locutor, y reconocía palabras discretas. Gana el U.S. National Award en 1972. Años después en Estados Unidos nace el interés de ARPA (Agencia de Proyectos de Investigación Avanzada) del U.S. Department of Defense, por el reconocimiento del habla continua, de vocabulario extenso. Los investigadores se enfocan su trabajo a procesos para el entendimiento del habla.

Los sistemas empiezan a incorporar módulos de:

- análisis léxico (conocimiento léxico)
- análisis sintáctico (Estructura de Palabras)
- análisis semántico (Significado)
- análisis pragmático (Intención)

El proyecto termina en 1976 con el resultado de que CMU, SRI, MIT crearon sistemas para el proyecto ARPA SUR (Speech Understanding Research).

Desde los años 80 a los 90 surgen sistemas de vocabulario amplio, que ahora son la norma (más de 1000 palabras). Empresas de los sectores de las comunicaciones y la informática trabajan en productos propios y comienzan las aplicaciones para ordenador personal.

En 1992 los laboratorios AT&T introdujeron su *Voice Recognition Call Processing System* aplicado a centralitas telefónicas. El sistema, a finales de 1993 procesaba 50 millones de llamadas al mes.

En 1995 aparecen los primeros teléfonos móviles que permitían los servicios de marcación por voz. En el mismo año comienza el desarrollo de los primeros prototipos de electrodomésticos controlados por voz (Whirlpool Corp.). Microsoft Corporation incluye facilidades para construir objetos de comandos de voz (voice-command objects) en su sistema operativo Windows 95

A partir del año 2.000 las centralitas de atención al cliente, incluyen el reconocimiento de voz basados en menú fijo de comandos de voz y reconocimiento de números para posteriormente integrar aplicaciones de diccionarios extensos. Creative Labs integra el procesamiento automático en la mayoría de sus tarjetas de sonido modelo soundblaster. Compaq y Pure Speech desarrollan conjuntamente tecnología de voz. Seagate Tech compró 25% de Dragon Systems. IBM desarrolla software de reconociendo de voz.

En la actualidad los dispositivos navegadores GPS más modernos incorporan una utilidad para entrada de información mediante voz y salida mediante síntesis, facilitando notablemente su manejo.

Como puede apreciarse, el crecimiento de los sistemas de reconocimiento de voz está en aumento y según los tiempos de vida que establecen las teorías de marketing sobre las tecnologías o productos, se puede decir que actualmente estamos en la primera fase, también llamada de “early adopters” en lo que se refiere a aplicaciones residenciales.

2.2 DESCRIPCIÓN DE SISTEMAS, ELEMENTOS Y NOMENCLATURA

Un aspecto crucial en el diseño de un sistema de reconocimiento de voz es la elección del tipo de aprendizaje a utilizar para construir las diversas fuentes de conocimiento.

Básicamente, existen dos tipos:

- **Aprendizaje Deductivo:** Estas técnicas se basan en la transferencia de los conocimientos que un experto humano posee a un sistema informático. Se llama también lenguaje basado en la explicación. Los inconvenientes son que exige un alto grado de conocimiento previo.

- **Aprendizaje Inductivo**: Estas técnicas se basan en que el sistema pueda, automáticamente, conseguir los conocimientos necesarios a partir de ejemplos reales sobre la tarea que se desea modelar. Requiere un número alto de ejemplos para el entrenamiento del sistema para ofrecer salidas o conclusiones con suficiente fundamento estadístico. En este segundo tipo, los ejemplos los constituyen aquellas partes de los sistemas basados en los modelos ocultos de Markov o en redes neuronales artificiales que son configuradas automáticamente a partir de muestras de aprendizaje.

En la práctica, no existen metodologías que estén basadas únicamente en el Aprendizaje Inductivo, de hecho, se asume un compromiso deductivo-inductivo en el que los aspectos generales se suministran deductivamente y la caracterización de la variabilidad, inductivamente.

2.2.1 Decodificador acústico-fonético

Las fuentes de información acústica, fonética, fonológica y léxica, con los correspondientes procedimientos interpretativos, dan lugar a un módulo conocido como decodificador acústico-fonético (o en ocasiones, decodificador léxico). La entrada al decodificador acústico-fonético es la señal vocal convenientemente representada; para ello, es necesario que ésta sufra un proceso previo de parametrización donde se determinan las características acústicas representativas de dicha señal. En esta etapa previa es necesario asumir algún modelo físico, contándose con modelos auditivos y modelos articulatorios. Una vez detectadas se etiquetan las unidades acústicas generándose cadenas de fonemas que se mapean con un vocabulario conocido para pasar a una estrategia de reconocimiento en función de las restricciones impuestas al reconocedor. Se indica a continuación en forma de diagrama de bloques en la figura 2.

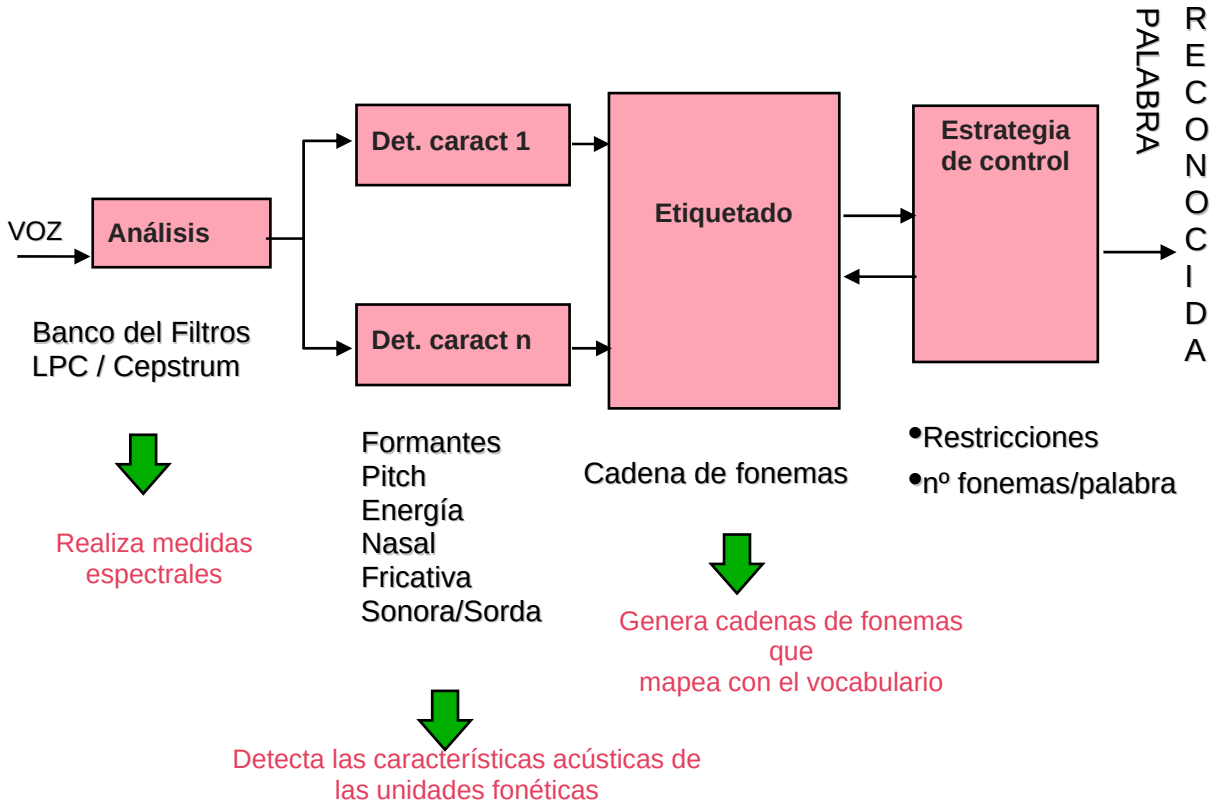


FIGURA 2

CONSIDERACIONES:

- Este método requiere un gran conocimiento acústico de las unidades fonéticas.
- El conjunto de características no es óptimo ya que se elige mediante la intuición.
- El diseño de clasificadores de sonido no es óptimo.

CONCLUSIÓN:

➢ No se usa en la práctica.

2.2.2 Sistemas basados en patrones

Es una base de datos de mútiples grabaciones de voz. Cada grabación es etiquetada y asociada a la información transcrita en modo texto cada palabra o frase y también asociada a los símbolos fonéticos que la representan. Decodifican lo pronunciado a partir de un conjunto de modelos (acústicos, lenguaje) que se captan de forma automática en una fase de entrenamiento, a diferencia del enfoque acústico-fonético que analiza directamente la voz para extraer las reglas que gobiernan el lenguaje.

Por tanto necesitan de un entrenamiento en una fase anterior al reconocimiento.

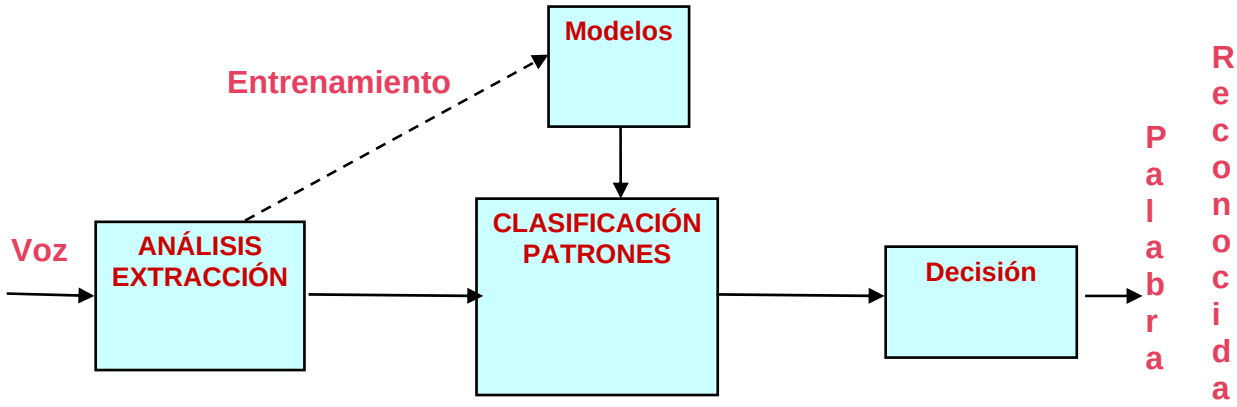


FIGURA 3

En fonética, la notación de la pronunciación de una lengua se hace según dos maneras normalizadas: el IPA y el SAMPA. Estas dos notaciones son comunes al conjunto de los fonetistas.

El **IPA** (International Phonetic Alphabet) representa cada fonema por un símbolo. Estos no pueden ser introducidos con un teclado de ordenador.

El **SAMPA** (Speech Assessment Methods Phonetic Alphabet) es una notación derivada del IPA, que puede ser tecleada. Los símbolos utilizados son símbolos ASCII.

CONSIDERACIONES:

- Requiere entrenamiento previo para disponer de una base de datos donde almacenar los modelos.

CONCLUSIÓN:

➢ En sistemas portátiles la base de datos debe ser pequeña para no incrementar costes de almacenamiento.

2.2.3 Modelo del lenguaje

Las fuentes de conocimiento sintáctico, semántico y pragmático dan lugar al modelo del lenguaje del sistema. Cuando la representación de la Sintaxis y de la Semántica tiende a integrarse, se desarrollan sistemas de reconocimiento de gramática restringida para tareas concretas.

2.3 RECONOCIMIENTO DE GRAMÁTICA RESTRINGIDA

El reconocimiento de la gramática restringida trabaja reduciendo las frases reconocidas a un tamaño más pequeño que la gramática formal. Este tipo de reconocimiento trabaja mejor cuando el hablante proporciona respuestas breves a cuestiones o preguntas específicas: las preguntas de “si” o “no”, al elegir una opción del menú, un artículo de una lista determinada, etc. La gramática especifica las palabras y frases más típicas que una persona diría como respuesta rápida y después asocia esas palabras o frases a un concepto semántico. Por ejemplo, un “si” puede entenderse cuando se oye un “sip”, “vale”, “yes” o “okey”, y un “no” con un “nop”, “nada” o “en absoluto”.

Estos sistemas se están utilizando en aplicaciones de servicios telefónicos de soporte y gestión virtuales. Si el hablante dice algo que gramaticalmente no tiene sentido, el reconocimiento fallará. Normalmente, si el reconocimiento falla, la aplicación incitará al usuario a repetir lo que ha dicho y el reconocimiento se intentará de nuevo. Si el sistema está correctamente diseñado y es repetidamente incapaz de entender al usuario (debido a que no ha entendido bien la pregunta, o por un acento cerrado, por interferencias o demasiado ruido alrededor), desviará la llamada a otro operador. La investigación muestra que las llamadas a las que se las pide replantear la pregunta o cuestión una y otra vez, en poco tiempo se frustran y cuelgan. Los sistemas que definen y resuelven el diálogo entre la máquina y el hombre de manera natural se llaman **gestores de diálogo.**

Los Interfaces conversacionales gestores de diálogo pueden ser de tres tipos:

- **SI** (System Initiative) en el que el sistema lleva la iniciativa en la conversación, acotando las posibilidades de respuesta.
- **UI** (User Initiative). El usuario lleva la iniciativa y el sistema se comporta de forma pasiva pidiendo aclaraciones.
- **MI** (Mixed Initiative). Una mezcla de los dos anteriores en los que la iniciativa puede ser llevada tanto por el sistema como por el usuario. Este tipo de sistema resulta mucho más natural pero también más complicado de implementar.

Los modelos del lenguaje complejos necesitan para su correcto funcionamiento grandes *corpus* de voz (bases de datos) y de texto escrito para la etapa de entrenamiento y aprendizaje. Gracias a ellos, se pueden abordar gramáticas más complejas y acercarse al. procesamiento del lenguaje natural.

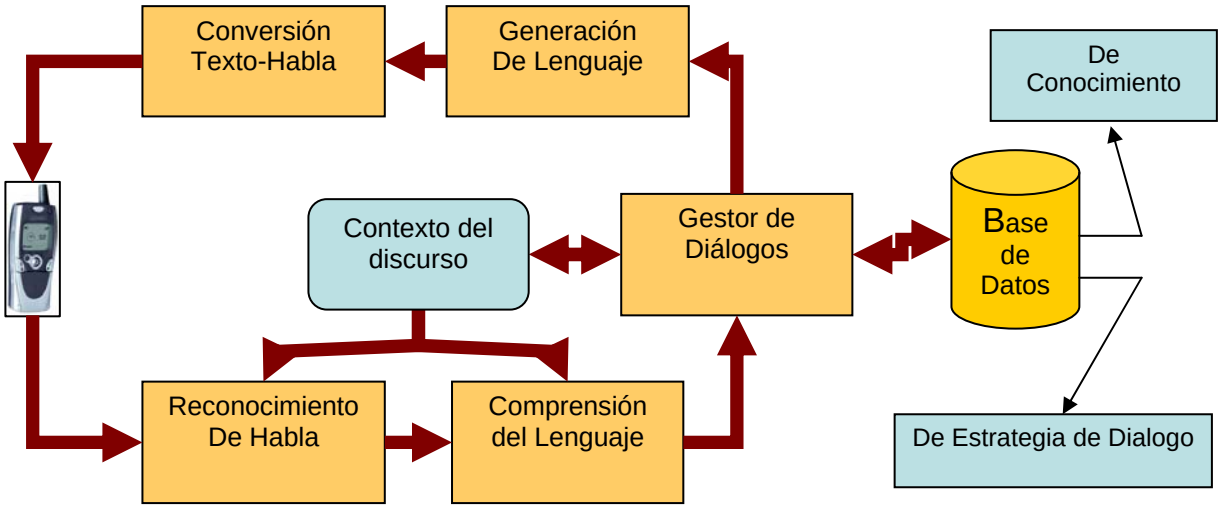


FIGURA 4

Se muestra como ejemplo en la figura 4 una aplicación de reconocimiento de voz con gestión de diálogo y síntesis de voz en respuesta a consultas telefónicas. Se comprueba que son varias las técnicas involucradas en el proceso. Por un lado la entrada vocal del usuario debe pasar por un reconocedor del habla que pasará la información al procesador lingüístico para la comprensión de la entrada. Posteriormente el gestor de diálogo hallará una estrategia de respuesta en función de la información disponible en la base de datos tanto de conocimiento como de estrategia de diálogo. Por último se generará una respuesta que será convertida de texto a voz.

2.4 PROCESAMIENTO DEL LENGUAJE NATURAL

El lenguaje natural es inherentemente ambiguo a diferentes niveles:

- A nivel léxico, una misma palabra puede tener varios significados, y la selección del apropiado se debe deducir a partir del contexto oracional o conocimiento básico. Muchas investigaciones en el campo del procesamiento de lenguajes naturales han estudiado métodos de resolver las ambigüedades léxicas mediante diccionarios, gramáticas, bases de conocimiento y correlaciones estadísticas.

- A nivel referencial, la resolución de anáforas y catáforas implica determinar la entidad lingüística previa o posterior a que hacen referencia.

- A nivel estructural, se requiere de la semántica para resolver la ambigüedad y la dependencia de los sintagmas preposicionales que conducen a la construcción de distintos árboles sintácticos. Por ejemplo, en la frase *Rompió el dibujo de un ataque de nervios*.

- A nivel pragmático, una oración, a menudo, no significa lo que realmente se está diciendo. Elementos tales como la ironía tienen un papel importante en la interpretación del mensaje.

- En la lengua hablada no se suelen hacer pausas entre palabra y palabra. El lugar en el que se debe separar las palabras suele depender del contexto o del énfasis que se le quiere dar al conjunto de la frase.

- El acento, bien de personas extranjeras, o simplemente diferentes formas de hablar dentro de una misma lengua así como dificultades en la producción del habla o expresiones no gramaticales hacen difícil la correcta interpretación de la frase.

- A nivel fonético. Muchas frases son muy parecidas fonéticamente y solo dentro de su contexto se resuelven correctamente. Como curiosidad señalar que los investigadores del grupo de reconocimiento de voz de Apple solían llevar una camiseta en la que se podía leer "*I helped Apple wreck a nice beach"* (ayudé a Apple a estropear una buena playa), cuya pronunciación es muy parecida a "*I helped Apple recognize speech* "(ayudé a Apple a reconocer el habla). Esta broma ilustra la dificultad de resolver correctamente cadenas fonéticas.

CONCLUSIÓN:

➢ Un sistema de lenguaje natural debe ser capaz de imitar algunos fenómenos típicamente humanos: confirmaciones durante la conversación, inicio de conversación menos fluido. Debe tener en cuenta que los diálogos entre humanos son variables con interrupciones frecuentes, solapamientos o frases incompletas o no estructuradas correctamente. La interacción con la máquina debe ser estructurada para que los objetivos del gestor de diálogo se realicen correctamente en corto espacio de tiempo.

2.5 CLASIFICACIÓN DE SISTEMAS DE RECONOCIMIENTO

Los sistemas de reconocimiento de voz pueden clasificarse según los siguientes criterios:

- **Entrenabilidad**: determina si el sistema necesita un entrenamiento previo antes de empezar a usarse.

- **Dependencia del hablante**: determina si el sistema debe entrenarse para cada usuario o es independiente del hablante.

- **Continuidad**: determina si el sistema puede reconocer habla contínua o el usuario debe hacer pausas entre palabra y palabra.

- **Robustez**: determina si el sistema está diseñado para usarse con señales poco ruidosas o, por el contrario, puede funcionar aceptablemente en condiciones ruidosas, ya sea ruido de fondo, ruido procedente del canal o la presencia de voces de otras personas.

- **Tamaño del dominio**: determina si el sistema está diseñado para reconocer lenguaje de un dominio reducido (unos cientos de palabras p. e. reservas de vuelos o peticiones de información meteorológica) o extenso (miles de palabras).

2.6 USO DEL RECONOCIMIENTO DE VOZ

Aunque en teoría cualquier tarea en la que se interactúe con un ordenador puede utilizar el reconocimiento de voz, actualmente las siguientes aplicaciones son las más comunes:

- **Dictado automático**: El dictado automático es el uso más común de las tecnologías de reconocimiento de voz. En algunos casos, como en el dictado de recetas médicas y diagnósticos o el dictado de textos legales, se usan corpus especiales para incrementar la precisión del sistema.

- **Control por comandos**: Los sistemas de reconocimiento de habla diseñados para dar órdenes a un computador (por ejemplo "Abrir Firefox", "cerrar ventana") se llaman Control por comandos. Estos sistemas reconocen un vocabulario muy reducido, lo que incrementa su rendimiento. Un caso a señalar que está teniendo un éxito y unas ventas espectaculares es el de la consola DS de Nintendo que incorpora programas educativos y de juegos que incorporan el control por voz como una entrada más de la aplicación.

- **Telefonía**: Algunos sistemas permiten a los usuarios ejecutar comandos mediante el habla, en lugar de pulsar tonos. En muchos casos se pide al usuario que diga un número para navegar un menú o palabras del propio menú.

- **Sistemas portátiles**: Los sistemas portátiles de pequeño tamaño, como los relojes o los teléfonos móviles, tienen unas restricciones muy concretas de tamaño y forma, así que el habla es una solución natural para introducir datos en estos dispositivos.

- **Sistemas diseñados para discapacitados**: Los sistemas de reconocimiento de voz se están utilizando para personas con discapacidades que les impiden teclear con fluidez. También hay experiencias de personas con problemas auditivos, que pueden usarlos para obtener texto escrito a partir de habla. Esto permite, por ejemplo, que los aquejados de sordera puedan recibir llamadas telefónicas.

- **Sistemas de control en automóviles**: Permiten mediante un conjunto de comandos reducidos controlar ciertos dispositivos del automóvil: limpiaparabrisas, intermitentes, etc. También se está utilizando el control por voz en navegadores GPS, básicamente para introducir los datos de destino: calles y números.

- **Sistemas de control del Hogar Digital.** Estamos en una primera fase en el que los sistemas de control de Hogar Digital están introduciendo el control por voz como un interfaz más de control del sistema. La mayoría de las aplicaciones son módulos adicionales que se pueden incluir para el control sobre protocolos de comunicaciones usados en el Hogar Digital. El sistema de la casa comercial Fagor, denominado “Maior Domo” dispone de una pulsera que reconoce los comandos de voz de forma independiente

del hablante y transmite las órdenes de forma inalámbrica a un dispositivo conectado a la red domótica.

2.7 SÍNTESIS DE VOZ

Hasta ahora se ha tratado el tema del reconocimiento automático de voz. Seguramente el lector automáticamente ha inferido que el sistema era capaz de contestar al usuario mediante voz. Para el hombre es natural el que la comunicación sea en los dos sentidos. El desarrollo de técnicas de síntesis de voz está evolucionando rápidamente con bastante éxito y su implantación definitiva es un hecho en ciertos sectores: por ejemplo, en atención telefónica automática y en el guiado por voz de navegadores GPS.

En el resto de campos potenciales de aplicación, entre los que incluimos el Hogar Digital existen unos condicionantes a tener en cuenta:

- En general, los usuarios no perdonan la falta de *naturalidad* y cierto timbre robótico, caracterizado por inflexiones del tono, poco naturales y voz con resonancia metálica.

- En lenguaje continuo o natural existe cierta reticencia según el sector de población analizado, a hablar con una “máquina”.

- No válido para aplicaciones que requieren privacidad.

Como ventajas evidentes, combinado con el reconocimiento de voz, cabe destacar su potencial como interfaz totalmente humanizado que minimiza el tiempo de aprendizaje del usuario.

Dentro del campo del lenguaje natural se están desarrollando aplicaciones de gestión de diálogo inteligente que permiten conversar de forma normal con el sistema, siendo este capaz de extraer información mediante análisis sintáctico y semántico y relacionando el contexto de la conversación. De esta forma la máquina se comporta de una forma mucho más humana. A su vez la síntesis del lenguaje emplea esa información para aplicar modulaciones en el tono y en la forma de hablar para acercarse al modelo de lenguaje verbal utilizado normalmente.

Hasta hace poco tiempo la mayoría de los sintetizadores de voz producían voces masculinas. Esto tiene una explicación técnica: es más fácil y ofrece mayor calidad la voz masculina que la femenina por tener la frecuencia del primer armónico más baja. Además la voz femenina tiene un componente de oscilación no periódico que viene dado por una mayor frecuencia en la aspiración, y que resulta más notable que la del hombre. Este componente de la excitación glotal es más difícil de modelar adecuadamente y al final resulta una voz menos real.

En la actualidad, sintetizadores de voz como Loquendo consiguen voces ”muy reales“ que simulan emociones y se acercan cada vez más a la forma de hablar humana. En el siguiente link el lector puede escribir o modificar el texto existente y el sintetizador de voz

online pasará a un fichero de audio que se podrá descargar al propio ordenador. Se puede comprobar como se incluyen expresiones y emociones.

Text To Speech Loquendo Demo Interactiva.

2.8 SISTEMAS DE RECONOCIMIENTO DE VOZ EXISTENTES

De los sistemas existentes en el mercado se incluyen aquellos más representativos por su presencia en el mercado, su integración con el hardware actual y sus cualidades:

a) Software comercial para ordenadores personales:

- Dragon Naturally Speaking de Nuance
- Philips FreeSpeech
- Protitle Live from NINSIGHT
- Via Voice de IBM
- Soluciones Loquendo
- Voice Pro 11 de Linguatec

b) Sistemas telefónicos.

- Nuance 8.5
- Telefónica: Software vocal de Telefónica
- Telisma (teliSpeech).

c) Software libre para ordenadores personales.

- CVoiceControl Se graba la orden como entrenamiento.
- PerlBox Sin entrenamiento, pero en inglés.
- Sphinx, del Sphinx Group en Carnegie Mellon University
- Open Mind Speech, antiguamente FreeSpeech

Las aplicaciones para sistemas telefónicos se ejecutan sobre sistemas de gran potencia de procesado (funcionamiento de varios canales simultáneos) y se combinan con sistemas de síntesis de voz (text to speech). Permiten el reconocimiento independiente del hablante con un gran margen respecto a acentos y variaciones en la pronunciación. Los primeros

sistemas únicamente realizaban reconocimiento de comandos sobre un menú prefijado.. En estos casos se apoyan en el uso del teclado numérico del teléfono para la selección de opciones de menú. De esta forma el usuario puede decir la opción del menú o pulsar la tecla correspondiente. Utilizan una confirmación del reconocimiento por parte del usuario antes de proceder al desvío de una llamada o a un menú posterior.

Los programas que corren sobre ordenadores personales, tanto comerciales como software libre, se orientan hacia el usuario final. Permiten el dictado automático de documentos, el control del sistema operativo y la navegación por la web mediante órdenes de voz. Son independientes del locutor y permiten el entrenamiento previo del sistema para la adaptación al locutor mejorando la precisión o bien corregir los errores sobre la marcha, haciendo que el sistema aprenda de forma dinámica. .La mayoría de las aplicaciones realizan un análisis del texto traducido para ayudarse en el reconocimiento según el contexto y mejorar la eficiencia.

Los precios de los software comerciales para reconocimiento de voz con ordenadores personales varían entre los 100 y 200 €. Su penetración en el mercado está realizándose como aplicaciones para dictado de documentos en oficinas. En el hogar su penetración es baja, el mayor uso de esta tecnología se da en personas con minusvalías físicas. Cabe esperar un mayor uso de esta tecnología en poco tiempo debido a la incorporación de "serie" de un programa de reconocimiento de voz en el sistema operativo Windows Vista de Microsoft.

Ventajas:

- Rapidez de dictado.
- Comodidad

Inconvenientes:

- Entrenamiento previo
- Fallos en reconocimiento y corrección por parte del usuario.

3 TECNOLOGÍAS EN EL RECONOCIMIENTO DE VOZ

El reconocimiento de voz, utilizado como una interfaz entre el hombre y la máquina se divide en 3 partes:

- Pre-procesamiento: Convierte la entrada de voz a una señal que el reconocedor pueda procesar.
- Reconocimiento: Identifica lo que se dijo (traducción de señal a texto).
- Comunicación: Envía lo reconocido al sistema (Software/Hardware) que actuará en consecuencia.

FIGURA 5

Pueden existir aplicaciones, en las que el interfaz de voz está integrado con el software o hardware de la aplicación. Estas pueden guiar al reconocedor especificando las palabras o estructuras que el sistema puede utilizar. Este caso sería un reconocedor por comandos y sólo algunas palabras podrían ser reconocidas. Por ejemplo. Dentro del Hogar Digital se pueden establecer una serie de palabras fijas: persiana, toldo, estor, cortina, luz, lámpara, aplique, televisión, radio, mp3... etc.

En el caso de reconocimiento de lenguaje natural el sistema ya conoce de antemano las palabras puesto que ya “ha sido entrenado previamente” y dispone de un corpus de voz.

Conviene señalar, que en ambos casos, es necesario indicar al sistema, de qué forma nos vamos a referir a los elementos del hogar. Por ejemplo, la lámpara del salón estará conectada a una toma de corriente controlada por el sistema de control energético. En

algún momento en el período de instalación y puesta en marcha del sistema es necesario asociar esa lámpara a dicha toma de corriente.

3.1 LOS DATOS EN EL RECONOCIMIENTO DE VOZ

Los sistemas de reconocimiento se basan en el análisis de los sonidos que distinguen una palabra de otra. Estos son los fonemas. Por ejemplo, "tapa", "capa", y "mapa", son palabras diferentes puesto que en su sonido inicial se reconocen fonemas diferentes
.

Existen varias maneras para analizar y describir el habla. Los enfoques más comúnmente usados son:

1. **Articulación:** Análisis de cómo el humano ***produce*** los sonidos del habla.
2. **Acústica**: Análisis de la ***señal*** de voz como una secuencia de sonidos.
3. **Percepción Auditiva:** Análisis de cómo el humano ***procesa*** el habla.

Los tres enfoques proveen ideas y herramientas para obtener mejores y más eficientes resultados en el reconocimiento.

La articulación da una información valiosa sobre la forma de producción de la voz. Queda fuera del ámbito del trabajo que nos ocupa aunque hay que señalar que se están realizando experimentos de reconocimiento de voz asociados al análisis y reconocimiento de imagen asociados a la articulación de tal forma que combinados, consiguen mejorar sustancialmente el número de aciertos en situaciones de ambigüedad o en casos con mala calidad de recepción acústica. Queda pendiente como futura línea de trabajo.

3.1.1 Acústica

Como se ha comentado en el punto anterior, un reconocedor acústico no puede analizar los movimientos en la boca. En su lugar, la fuente de información es la señal de voz misma.

El habla es una señal analógica, es decir, un flujo continuo de ondas sonoras y silencios. La acústica se utiliza para identificar y describir los atributos del habla que son necesarios para un reconocimiento de voz efectivo.

Las características importantes del análisis acústico son:

1. Frecuencia
2. Amplitud
3. Resonancia.
4. Formantes y espectrograma.

3.1.2 Frecuencia y Amplitud

Todos los sonidos causan movimientos entre las moléculas del aire. Algunos sonidos, tales como los que produce una cuerda de guitarra, producen patrones regulares y prolongados de movimiento del aire. Los patrones de sonidos más simples son los sonidos puros, que se pueden representar gráficamente por una onda sinusoidal. En la práctica los sonidos puros no existen. Matemáticamente es posible descomponer cualquier sonido compuesto en una sucesión de sumas de sonidos puros de diferentes frecuencias y amplitudes. Esta técnica nos da una herramienta de cálculo y computación muy potente que nos permite estudiar los sonidos tanto en el dominio del tiempo como en el dominio de la frecuencia.

Se puede definir la frecuencia como el número de vibraciones (ciclos) del tono por segundo. Se mide en hercios (Hz).

Tonos altos = Mayor frecuencia
Tonos bajos = Menor frecuencia

El volumen de un sonido refleja la cantidad de aire que es forzada a moverse.
Se describe y representa como amplitud de la onda y se mide en decibelios (dB).

3.1.3 Resonancia

La mayoría de los sonidos incluyendo del habla tienen una frecuencia dominante llamada frecuencia fundamental que la percibimos como el pitch (tono) combinado con frecuencias secundarias. En el habla, la frecuencia fundamental es la velocidad a la que vibran las cuerdas vocales al producir un fonema sonoro.

Sumadas a la frecuencia fundamental hay otras frecuencias que contribuyen al timbre del sonido. (Son las que nos permiten distinguir una trompeta de un violín, etc. o las voces de diferentes personas). Algunas bandas de la frecuencia secundarias juegan papel importante en la distinción de un fonema de otro. Se les llama formantes y son debidas a la resonancia.

La resonancia se define como la habilidad que tiene una fuente vibrante de sonido de causar que otro objeto vibre (por ejemplo en una fábrica, una máquina hace que vibre el piso). Las cámaras de resonancia en instrumentos de música responden a frecuencias específicas o anchos de banda específicos. Al ser estas cajas o cámaras de resonancia más grandes que la fuente del sonido amplifican las frecuencias a las que responden.

La garganta, boca y nariz son cámaras de resonancia que amplifican las bandas o frecuencias formantes contenidas en el sonido generado por las cuerdas vocales. Estas formantes amplificadas dependen del tamaño y forma de la boca y si el aire pasa o no por la nariz.

La resonancia que caracteriza el timbre de una vocal resulta del filtrado que sufre la vibración de las cuerdas vocales al pasar por la boca comportándose como un filtro. Las frecuencias que la boca deja pasar son diferentes para cada vocal, principalmente porque las cavidades de resonancia que las filtran cambian de forma y de dimensiones. Estas formas o dimensiones son diferentes en cada individuo y son los que definen el timbre de cada persona. Esta característica define una huella biométrica que permite el **reconocimiento del hablante**, siendo útil en sistemas de comprobación de identidad o para definir perfiles de uso diferentes según el usuario que habla.

3.1.4 Formantes y Espectrograma

El habla humana es una combinación de múltiples frecuencias e intensidades y se representa como una onda compleja.

Las vocales se pueden descomponen en dos o más ondas periódicas simples. También al hablar, se emiten ondas no periódicas que forman parte de todos los fonemas sonoros, consonantes y semivocales. Las frecuencias y características de los patrones no periódicos proveen información importante sobre la identidad de los fonemas.

Cada fonema tiene un patrón único de energía a diferentes frecuencias. Analizando estos picos de intensidad (energía) en el espectro, es decir en el dominio de la frecuencia, se pueden distinguir las componentes del habla humana. Estos picos de frecuencias en el espectro es lo que llamamos formantes. La herramienta de análisis, utilizando técnicas como la transformada rápida de Fourier (FFT), es el espectrograma.

La identidad de las consonantes se revela por el cambio en las formantes que resultan cuando los articuladores se mueven de un fonema anterior a la consonante y de ella al siguiente fonema llamadas transiciones de formantes.

3.2 CODIFICACIÓN DE LAS SEÑALES

Para utilizar la voz como un dato utilizable en aplicaciones digitales es necesario seguir una serie de pasos:

El habla es una señal continua que varia en el tiempo. Las variaciones en la presión del aire se irradian desde la cabeza y se transmiten por el aire. Un micrófono convierte esas variaciones en presión del aire a variaciones en voltaje convirtiéndolas en una señal analógica.

3.2.1 Muestreo

Es necesario digitalizar la señal y obtener una representación binaria fiel a la señal original. Normalmente las señales digitales ser codifican usando la técnica PCM (Pulse Code Modulation).

Tomando a intervalos de tiempo fijos una muestra de la señal y asegurando que la frecuencia de muestreo (inversa de los tiempos de muestra) es por lo menos el doble de la frecuencia máxima de la señal a capturar, aseguramos que la representación de la señal y su posterior reconstrucción son correctas. Esto se describe en el teorema de Nyquist-Shannon.

Para poder reconstruir la señal a partir de las muestras, es necesario debe pasar por un filtro *paso-bajo* a la frecuencia de muestreo.

Si no se cumple lo anteriormente expuesto se produce el denominado efecto de *aliasing*.

El humano produce señales de Voz desde los 100 Hz (hombre) o los 400 Hz (mujer) hasta los 15.000 Hz. Ejemplos:

Teléfono: 3100 Hz de ancho de banda y frecuencia de muestreo a 8000 Hz, inteligible pero con baja calidad.

CD de audio, 20 Khz. de ancho de banda con muestreo a 44,1 Khz.

DVD audio, 20 Khz. de ancho de banda con muestreo a 48 Khz.

El ancho de banda es mayor para instrumentos que para voz. Por lo tanto se requiere mayor espacio para almacenar y transmitirla.

En el caso que nos ocupa, aunque se ha indicado que el habla humana puede generar frecuencias de hasta 15.000 Hz, en la realidad esas altas frecuencias no contienen una información necesaria para la correcta interpretación. .Se puede considerar que un ancho de banda de 8.000 Hz es suficiente para una representación muy aproximada de los mensajes sonoros, por tanto, según el teorema de Nyquist, la frecuencia de muestreo mínima para aplicaciones de reconocimiento de voz será de 16.000 Hz. A mayor frecuencia de muestreo, mayor detalle de la señal y también mayor tamaño del fichero de audio.

CONCLUSIÓN:

- Valor óptimo de frecuencia de muestreo: 16.000 Hz

3.2.2 Resolución: Codificación

Cada muestra se representa con un valor digital limitando el rango de valores discretos correspondiente al original. Utilizando n bits se pueden representar 2^n valores diferentes. (con 8 bits , 256 posibles valores). A mayor número de bits mayor número de valores y más pequeño es el escalón entre dos posibles valores. Este concepto se define como

resolución. En el caso de un CD de audio la resolución del proceso de cuantificación es de 16 bits pudiendo tener 65.536 posibles valores de amplitud de la señal.

El error o diferencia entre la señal original y la reconstruida de llama ruido de cuantificación. Por lo tanto, a mayor resolución menor ruido. Como contrapartida una codificación de la señal con mayor número de bits influye en la carga computacional a la hora de trabajar con dichas señales.

La resolución del codificador (de B bits por muestra) se describe generalmente en términos de la relación señal-ruido (SNR).

A mayor SNR, mayor la fidelidad de la señal digitalizada. Señalar que es independiente de la frecuencia de muestreo.

$$SNR = 2^{B}$$

Donde B=bits / muestra

Ejemplo:

Teléfono: 8 bits por muestra, es decir, si muestreamos a 8 Khz. tenemos 8.000 muestras por segundo; 8000 x 8 = 64.000 bits por segundo.

CD: 16 bits por muestra, por lo tanto 44.100 muestras por segundo x 16 bits = 705.600 bits por segundo (mono).

Para 2 canales (estéreo) se duplica la tasa siendo el total de: 1.411.200 bits por segundo. Esta tasa está muy por encima de las necesidades en el procesado de voz, por un lado se puede eliminar un canal y por otro, se considera que una representación fiel para el procesado de una señal de voz se consigue con al menos 10 bits, disponiendo de 1.024 posibles niveles.

CONCLUSIÓN:

- La señal de voz puede codificarse con 10 bits manteniendo una calidad suficiente para su análisis y reconocimiento.

3.3 DESCRIPCIÓN DE TÉCNICAS DE RECONOCIMIENTO ACTUALES

Hay que distinguir entre los sistemas de reconocimiento de habla automáticos, que intentan transcribir en lenguaje escrito lo que un locutor ha expresado oralmente, de los sistemas de comprensión del lenguaje. Estos últimos sistemas tratan un concepto más amplio, que pudiendo incluir entre otras partes un sistema de reconocimiento de habla automático, su objetivo final es la semántica del mensaje, es decir, comprenderlo.

3.3.1 Dynamic Time Warp (DTW)

Básicamente es un algoritmo que es capaz de medir la similitud entre dos secuencias (señales de audio) entre las cuales puede haber una variación en la velocidad o en el tiempo. Se almacenan previamente patrones de voz típicos como modelos de referencia en un diccionario de palabras candidato. El reconocimiento se lleva a cabo comparando la expresión desconocida con los patrones almacenados y se selecciona el que más se le parece. Normalmente se almacenan patrones para palabras completas.

Ventaja:

- Se evitan errores debidos a la segmentación o clasificación de unidades pequeñas que puedan variar mucho acústicamente, como los fonemas.

Inconveniente:

- Cada palabra requiere de un patrón previo almacenado. El tiempo de cálculo puede ser muy alto si el número de palabras posible (vocabulario) es grande. A su vez el tiempo de entrenamiento es alto en comparación con otros sistemas.

Inicialmente se pensaba que se restringía sólo a reconocimiento dependiente del locutor. Sin embargo, es posible un reconocimiento independiente del locutor. Esto se consigue utilizando técnicas de clustering para generar automáticamente grupos de patrones para cada palabra del vocabulario.

Se considera esta técnica del tipo determinista pues explora características conocidas de la señal. Es eficaz para sistemas de reconocimiento de habla automáticos de vocabulario restringido.

3.3.2 Modelos ocultos de Markov (Hidden Markov Models)

Es un modelo estadístico cuyo objetivo es determinar los parámetros ocultos o desconocidos de una cadena de la cual se conocen, o son observables, ciertos parámetros. Con estos parámetros extraídos se pueden realizar sucesivos análisis. Se caracteriza por tener estados. El sistema evoluciona de unos estados a otros con transiciones probabilísticas. El estado no es visible directamente sino que sólo lo son las variables influidas por el estado. Cada estado tiene una distribución de probabilidad sobre

los posibles símbolos de salida. El estado en t+1 solo depende del estado en t y no de la evolución anterior del sistema. En consecuencia la secuencia de símbolos generada por un HMM proporciona información sobre la secuencia de estados.

Ventaja:

- Está indicado para reconocimiento de lenguaje natural o continuo.

Inconveniente:

- Es necesario un entrenamiento previo sobre el llamado corpus de voz. Además el procesado, en comparación con el reconocimiento del habla con vocabulario restringido, requiere sistemas potentes para conseguir tiempos bajos de respuesta.

Para comprender como se aplica esta técnica podemos decir que para cada texto el locutor da una emisión sonora que es convertida posteriormente por un procesado en una señal digital. Imaginemos que para cada una de estas posibles emisiones hemos encontrado un modelo capaz de imitar al locutor, es decir, el modelo es capaz de regenerar la misma señal que el locutor. Así suponemos que tenemos tantos modelos como posibles emisiones y para cada modelo, hay por supuesto, un texto asociado. Teniendo entonces una determinada emisión del locutor podemos encontrar de entre todos los modelos que genera dicha emisión y su texto asociado. Es una aplicación de los modelos un tanto peculiar puesto que normalmente se construyen los modelos que ante unas entradas determinadas genera unas salidas en principio no conocidas. En este caso se realiza una búsqueda del modelo exacto que puede dar una determinada y conocida salida.

Si los modelos necesarios, en principio infinitos, no son independientes entre sí y teniendo en cuenta que el habla tiene una organización estructural jerárquica (fonemas, palabras, frases), es posible construir una gran cantidad de modelos combinando un número razonable de pequeñas partes. Esto reduce significativamente el tamaño de la base de datos. Existen algoritmos que permiten buscar de forma probabilística, ahorrando tiempo de cálculo, el modelo que más se parece a la cadena buscada, por ejemplo, algoritmo de Viterbi.

Existe una herramienta de libre uso desarrollada por el departamento de ingeniería de la Universidad de Cambridge llamado HTK (Hidden Markov Model ToolKit) que dispone de documentación y ejemplos para uso académico y de estudio.

3.3.3 Redes Neuronales

Este tipo de sistemas se basan en el entrenamiento de una red neuronal. Clasifican los fonemas según sus características concretas: energía en función del tiempo o de la frecuencia y ancho de banda, para la extracción de rasgos, obteniendo vectores con pesos determinados. La red, como todas las redes neuronales, necesita de un entrenamiento previo de forma iterativa para ir ajustando los pesos. La salida de la red es comparada con la salida esperada (y conocida de antemano) calculándose un error. En la actualidad es un

campo que está en constante desarrollo y cuyas aplicaciones están cada vez más justificadas dentro de los campos de reconocimiento de voz, de imagen o de patrones.

Ventaja:

- Requiere de una potencia de cálculo menor que los modelos ocultos de Markov ante un corpus de voz similar.

Inconveniente:

- Para un mismo corpus de voz, en comparación con los modelos ocultos de Markov el tiempo de entrenamiento es mayor.

En la actualidad se están utilizando combinaciones de redes neuronales y modelos ocultos de Markov para conseguir una mejora en el procesado y en el reconocimiento. Un producto que emplea ambas técnicas es Loquendo ASR.

3.4 INTERFACES DE USUARIO

Se define el interfaz de usuario como el conjunto de componentes empleados por los usuarios para comunicarse con sistemas electrónicos u ordenadores. El usuario dirige el funcionamiento mediante entradas. Las entradas se pueden generar mediante diversos dispositivos: un teclado, un ratón, un mando a distancia, un gesto o la voz. Estas entradas se interpretan y se convierten en señales electrónicas o eléctricas que pueden ser procesadas por el sistema. Una vez que se han ejecutado las instrucciones indicadas por el usuario el sistema puede comunicar los resultados mediante salidas que serán interpretadas por dispositivos de salida: una pantalla, una luz, un altavoz, etc.

En la práctica la mayoría de los interfaces que utilizamos son interfaces multimodales, es decir, que permiten su uso de formas diferentes. En el caso de un ordenador, se puede interaccionar con el sistema operativo mediante el teclado, mediante un ratón o mediante la voz. En ocasiones cierto interfaz será más indicado para realizar alguna acción o bien por preferencias del usuario le resultará más cómodo o intuitivo.

3.4.1 CARACTERÍSTICAS HUMANAS EN EL DISEÑO DE INTERFACES

Al diseñar interfaces de usuario deben tenerse en cuenta las habilidades cognitivas y de percepción de las personas, y adaptar el interfaz o la forma de actuar del sistema a ellas.

Una de las cosas más importantes que una interfaz debe conseguir es reducir la dependencia de las personas de su propia memoria no forzándoles a recordar cosas innecesariamente (por ejemplo, información que apareció en una pantalla anterior) o a repetir operaciones ya realizadas (por ejemplo, introducir un mismo dato repetidas veces). Además es necesario estudiar el tipo de usuario que va a manejar el interfaz ya que los

procesos cognitivos varían considerablemente en sectores diferentes de la población, así lo que para una persona acostumbrada a manejar tecnología puede resultar simple, para otra no acostumbrada puede suponer un serio problema.

Sintetizando, se muestran a continuación los parámetros más importantes a la hora de diseñar una interfaz.

- Velocidad de Aprendizaje.- Se pretende que la persona aprenda a usar el sistema lo más pronto posible.
- Velocidad de Respuesta.- El tiempo para realizar una operación en el sistema.
- Tasa de errores.- Porcentaje de errores que comete el usuario.
- Retención.- Cuánto recuerda el usuario sobre el uso del sistema en un período. de tiempo.
- Satisfacción.- Se refiere a que el usuario esté a gusto con el sistema.
- Características Físicas.- Cada persona tiene diferentes características físicas.
- Ambiente.- El lugar donde va a ser usado el sistema. Cada interfaz tiene que adecuarse al lugar.
- Personalidad.- De acuerdo a la edad, nivel socio-económico, etc.
- Cultura.- Los japoneses no tienen las mismas pantallas, ventanas, etc. Este factor es importante si el mercado para el sistema es a nivel internacional.

Los interfaces diseñados para su uso en el hogar deben tener en cuenta en el diseño pensado en la accesibilidad. Este concepto se llama *Diseño para Todos.* Esto significa que al diseñar un sistema, un servicio o un producto, se debería tener en cuenta que tiene que ser fácilmente utilizable para personas también con discapacidades físicas e intelectuales.

Se abre en este punto una posible línea de investigación como futuro trabajo en el estudio de interfaces de voz diseñados para el uso de personas con discapacidades físicas (problemas en el habla o en la audición) o discapacidades intelectuales.

3.4.2 EL DIÁLOGO INTELIGENTE

La idea fundamental que subyace a este tipo de sistemas es que el usuario puede tomar la iniciativa durante el diálogo, y, por tanto, no debe estar ceñido a un plan o menú previamente elaborado. El usuario debe poder cambiar de intención en cualquier momento del diálogo, sin tener previamente que volver 'al menú principal'. Además, el sistema debe ser colaborativo y cooperativo, involucrándose con el usuario en la consecución de una tarea común.

Un caso extremo del aspecto colaborativo es ser 'proactivo', esto es, tomar la iniciativa y proponer una acción ante una situación concreta, como puede ser un escape de gas o la activación de una alarma.

Idealmente, además, el sistema debe ser capaz de adaptarse a diversos usuarios en función de su nivel de conocimiento del sistema en sí mismo, de sus preferencias anteriores, o del contexto en el que se encuentre. Por último, el usuario, no tiene que 'aprender' una serie de comandos o expresiones fijas para dirigirse al sistema. Al contrario, debe ser capaz de dirigirse a él de forma natural, usando las expresiones propias del lenguaje natural, tal y como se dirigiría a una persona.

4 ANÁLISIS DE LA INTEGRACIÓN DEL CONTROL POR VOZ EN EL HOGAR DIGITAL

Una vez mostrado el abanico de posibilidades y aplicaciones del control por voz e indicado las pautas para el diseño de interfaces de usuario, nos centramos a continuación en la aplicación de estas técnicas en un entorno determinado. Este entorno es el Hogar Digital. Previamente es necesario definir el concepto de Hogar Digital.

El concepto del Hogar Digital en una vivienda se entiende como el conjunto de infraestructuras, equipos y sistemas que se integran tecnológicamente. Ofrece a sus habitantes funciones y servicios que facilitan su gestión y mantenimiento, aumentan la seguridad; incrementan el confort, mejoran las telecomunicaciones, ahorran energía, costes y tiempo. Además ofrece nuevas formas de entretenimiento, ocio y prepara al hogar para la integración de futuros servicios otros servicios dentro del mismo y de su entorno. En definitiva, se trata de mejorar la calidad de vida de las personas. Para ello existirán diferentes soluciones en función de las necesidades de cada familia o personas que habiten el Hogar Digital.

Tradicionalmente se ha asociado el uso de automatismos (domótica) con viviendas de alto nivel y como algo muy complejo y sofisticado. En la actualidad este concepto está cambiado rápidamente ya que los hogares cada día disponen de un mayor número de dispositivos de alta tecnología: televisores, ordenadores, reproductores multimedia, videoconsolas, electrodomésticos programables, sistemas de seguridad y telecomunicaciones.

En el día a día cada uno de estos dispositivos cumple perfectamente el objetivo para el que está diseñado y el grado de satisfacción de los usuarios con estos dispositivos suele ser alto. El problema surge cuando se comprueba que los dispositivos no se integran dentro del hogar y en la mayoría de los casos tampoco entre ellos. No hay más que echar un vistazo a la mayoría de los salones de los hogares para comprobar el número de mandos a distancia que suelen encontrarse. Lo ideal sería poder controlar todos los dispositivos con un solo mando a distancia. Existen mandos universales capaces de reconocer un amplio número de dispositivos y controlarlos. Las experiencias con estos

mandos es que, o bien son muy complicados de manejar, o no disponen de algunos de los códigos de control de algún dispositivo de audio o video y acaban desechándose. Lo mismo ocurre con los sistemas de calefacción o de aire acondicionado o con los electrodomésticos clásicos: nevera, lavadora, lavavajillas, cada uno dispone de un sistema de gestión que es necesario aprender a manejar.

Toda esta amalgama de sistemas, equipos y servicios se integra en el Hogar Digital, capaz de integrar diferentes sistemas de control y hacerlos transparentes al usuario controlándolos de una forma sencilla y cómoda.

El control y gestión de este hogar se está realizando en la actualidad mediante el uso de interfaces como pulsadores, interruptores, pantallas táctiles, PDAs y ordenadores portátiles. En algunos casos solo se instala un solo interfaz, mientras que en otros casos pueden existir diferentes interfaces para realizar una misma acción. Por ejemplo, se puede apagar una luz actuando sobre el interruptor situado en la pared, o de forma remota mediante el uso de un ordenador que se conecta de forma segura a nuestra vivienda a través de internet.

4.1 Usuarios del control por voz

El uso de la voz para el accionamiento de ciertos elementos del entorno puede interpretarse como una comodidad o un lujo si se analiza desde el punto de vista de personas sin problemas de dependencia.

Para aquellas personas con problemas de dependencia física: minusválidos o personas de la tercera edad con movilidad reducida, el hecho de poder controlar por si mismos el encendido o apagado de una luz o la subida o bajada de una persiana, supone una gran mejora de su calidad de vida. Este es el caso más justificable para la implementación de la tecnología de control por voz. De hecho es el sector que está impulsando el nacimiento de nuevas empresas dedicadas al Hogar Digital Accesible.

En cualquier caso, cualquier sistema que facilita o mejora aspectos relacionados con acciones cotidianas dentro del hogar es siempre percibido por los usuarios como algo positivo. Por tanto, este estudio analiza técnicamente el control por voz con unos criterios generales, respecto a los usuarios a fin de abordar todo el abanico de posibilidades que ofrece esta tecnología.

4.2 Tipologías de sistemas de control por voz

Se parte de la necesidad de contar en cada estancia con elementos captadores de voz además de elementos de respuesta mediante síntesis de voz. Los elementos captadores deben son micrófonos con pre-amplificadores para una correcta adaptación de las señales que deben llegar a la unidad central o de procesamiento. Los dispositivos de respuesta son altavoces de audio de baja potencia. Los circuitos pre-amplificadores asociados a los

micrófonos también deben disponer de alimentación para la polarización de los elementos activos del amplificador de audio.

Desde un punto de vista técnico, existe la posibilidad de realizar el procesado del reconocimiento de voz en el propio elemento captador situado en cada estancia. Desde el punto de vista económico, el hecho de introducir una capacidad de proceso elevada a muchos dispositivos (uno por cada estancia de la vivienda) encarece sustancialmente el precio final de todo el sistema. Por tanto, aunque parece haber alguna iniciativa en el mercado (que se comentará más adelante) en esta tipología, no se estudia técnicamente en este trabajo por considerarla fuera de mercado. Se considera una tipología en el que la inteligencia del sistema, es decir el procesado de las señales de audio procedentes de los elementos captadores (micrófonos) y la síntesis de voz que se transmitirá a los elementos emisores (altavoces), se realiza en una unidad de procesamiento central, o sistema central.

Si el transporte de las señales de audio se realiza de forma analógica, puede ser necesario amplificar la señal que llega a los altavoces desde el sistema central de gestión. Si es así, cada elemento de respuesta deberá contar con un circuito de amplificación integrado y será necesario alimentar eléctricamente el amplificador de audio para el altavoz. La otra posibilidad es que el sistema central de gestión disponga de módulos amplificadores de las señales de salida.

Queda como trabajo futuro el estudio económico de las dos posibilidades: la primera con amplificadores distribuidos, uno por cada altavoz; y la segunda, con amplificadores incluidos en las salidas del sistema central.

Es necesario examinar y valorar las diferentes tipologías posibles del sistema para, dependiendo de las necesidades primarias, elegir adecuadamente según el caso.

4.2.1 Conexionado en estrella de elementos captadores y emisores de voz

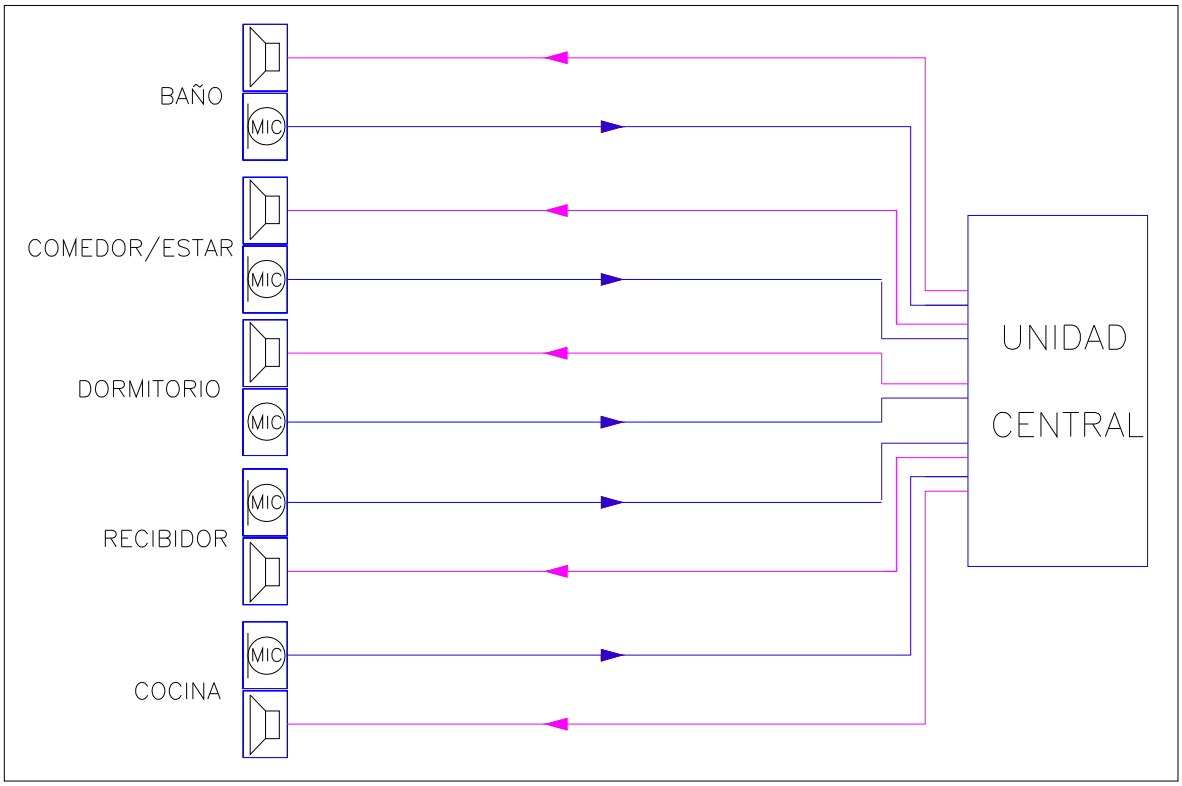


FIGURA 6

En la figura 6 se ilustra una topología de conexionado con cableado dedicado desde la unidad central hasta cada estancia en una topología en estrella.

Si las señales son analógicas, se recomienda que el conexionado se realice de forma balanceada para dotarlas al sistema de una mayor inmunidad al ruido eléctrico y mejorando la relación señal-ruido.

Si el transporte de las señales se realiza en formato digital se recomienda el uso del formato digital AES/EBU, utilizado en audio profesional en lugar del formato SPDIF, derivado del primero pero pensado para aplicaciones menos exigentes. La razón es la mayor inmunidad frente a descargas electrostáticas y conmutaciones, muy frecuentes en el hogar debido a electrodomésticos, iluminación y dispositivos electrónicos. En este caso la parte de amplificación de la señal de audio irá siempre en el hardware asociado al altavoz. Este hardware será el que realice la conversión D/A previa a la amplificación.

La unidad central debe disponer de tantas entradas (analógicas o digitales, según el caso) como micrófonos sea necesario instalar. De igual forma, dispondrá de tantas salidas como altavoces haya instalados.

Se ha considerado como ejemplo para mostrar en los esquemas, una vivienda con un solo dormitorio y en la que se instala el sistema de control en todas las estancias incluyendo baño y recibidor.

En viviendas con estancias de gran tamaño puede ser necesario instalar dos elementos captadores y dos altavoces en la misma estancia para asegurar tanto la captación como la escucha de los mensajes de salida. En ese caso pueden asociarse las señales que llegan a los altavoces dentro de una misma estancia. No podría realizarse la misma asociación con los micrófonos ya que podrían existir problemas de reducción de señal por la aparición de contra-fases en las señales.

CONCLUSIONES:

- Desde el punto de vista técnico, el desarrollo con señales analógicas es más simple y menos costoso. Si la instalación del cableado de audio se realiza correctamente y las distancias entre la unidad central y los dispositivos situados en las estancias, no superan el centenar de metros, es una solución válida y fiable.

- La utilización de señales digitales permite aumentar la longitud del cableado respecto a señales analógicas sin perdida de calidad. El coste es mayor ya que debe utilizar conversores A/D y D/A en el sistema de control y en los dispositivos hardware emisores respectivamente.

- Con esta tipología siempre deben existir dos cables, uno de ida, que transporta el audio desde el sistema central hasta los dispositivos emisores, y otro de vuelta, que transporta el audio desde los elementos captadores hasta el sistema central.

4.2.2 Conexionado mediante red LAN

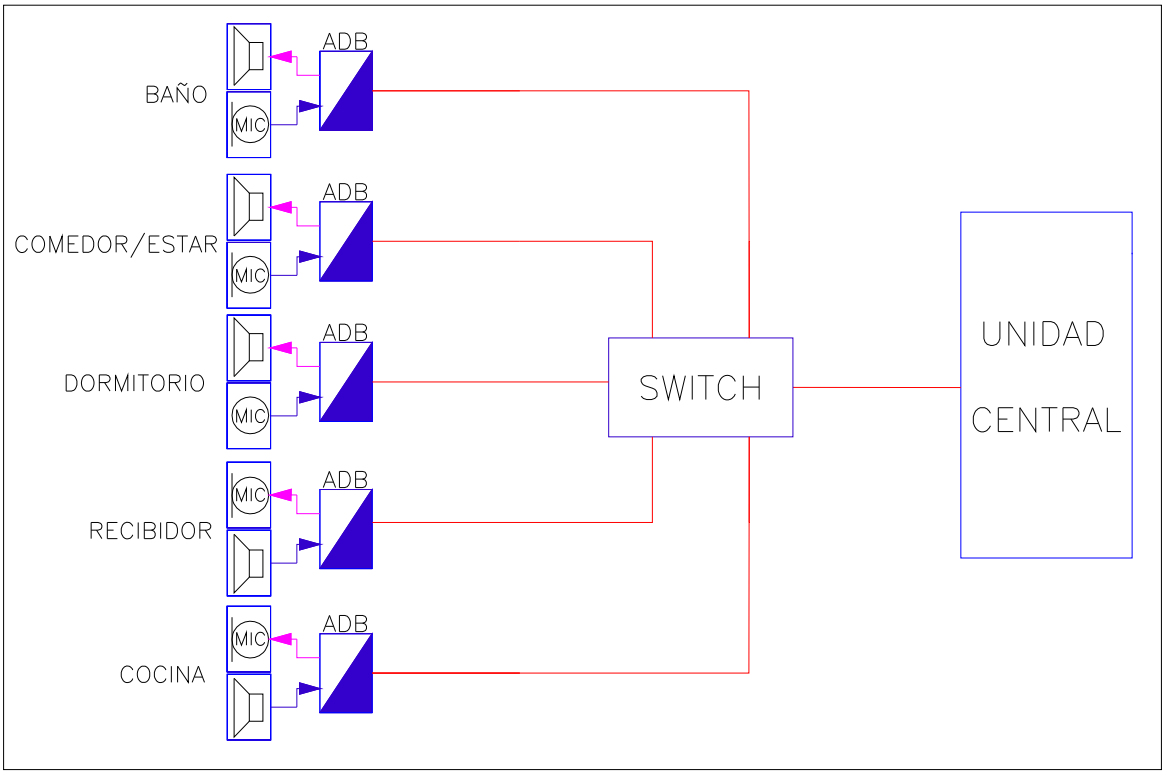


FIGURA 7

En la figura 7, la diferencia respecto a la figura 6 es que la conexión se realiza mediante una red LAN cableada. Para ello se ha introducido en cada estancia un adaptador de BUS (ADB) para permitir el flujo de señales en ambos sentidos entre los captadores y altavoces con la unidad central. La diferencia entre esta solución y la anterior radica en la posibilidad de utilización de un BUS. Para ello es necesario contar con elementos adaptadores de BUS que permiten el flujo de información en ambos sentidos. Desde el punto de vista de instalación en una vivienda es mucho mejor esta segunda opción por la utilización de un solo cable, que reduce el coste de instalación.
En el primer caso del punto 3.2.1 cada estancia necesita cableado doble y el número de entradas y salidas de audio necesarios en la unidad central es alto, incrementando su coste de fabricación.

Como contrapartida, en este segundo caso el incremento de coste deriva de la inclusión de los adaptadores de BUS que, además de contar con la electrónica necesaria para la conectividad a la red LAN, deben incluir un conversor analógico-digital (AD), para la captura de señales procedentes del micrófono, y un conversor digital-analógico para la conversión de las tramas digitales a audio analógico que debe reproducir el altavoz. Se ilustra a continuación un esquema con los elementos que debe contar el adaptador de BUS

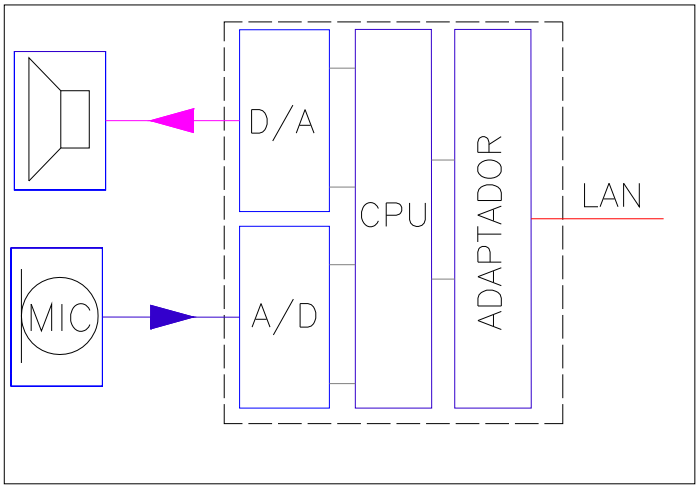


FIGURA 8

4.2.2.1 Cálculo del Bit Rate

Según lo indicado en puntos anteriores, el ancho de banda de la voz, para una representación correcta podemos estimarlo en 8.000 Hz. Según el teorema de Nyquist, la frecuencia de muestreo debe ser por lo menos el doble: 16 KHz. La resolución que asegura un detalle suficiente para el análisis, sin llegar a calidad CD (16 bits/ muestra) se fija en 10 bits por muestra. Por tanto:

- 10 bits/muestra x 16.000 muestras/s = **160 kbps** para cada canal de audio.

Hay que tener en cuenta que esta tasa es sólo para un canal. En el caso que nos ocupa, con 5 estancias y dos canales por estancia, 160 x10 = 1.600 kbps, la tasa aumenta considerablemente aunque sigue siendo admisible Esta tasa binaria es aceptable en relación a la capacidad de una red LAN a 100 Mbit/s. Para la parte de respuesta (síntesis de voz) se pueden estimar unas tasas similares por canal.

Para el caso analizado con 5 estancias, suponiendo una simultaneidad del 100%, el bit rate sería de 3.200 kbps.

Este cálculo hipotético no se dará en la práctica ya que el sistema o escucha o contesta pero nunca realizará las dos acciones simultáneamente. Una fórmula práctica para calcular la tasa binaria puede ser dividir entre dos el máximo calculado.

- 3.200 / 2 = 1.600 kbps = 1,6 MBit/s

Utilizando una red LAN a 100 MBit/s el bit rate es suficiente para asegurar una transmisión sin problemas permitiendo compartir el uso de la red con otras aplicaciones.

Una recomendación en el diseño de la red LAN es usar un SWITCH independiente (o una parte de el) para el tráfico de la red de reconocimiento y síntesis de voz. Esta separación siempre es recomendable cuando dentro de la misma red hay sistemas que utilizan la red con una dedicación mayor: streaming de audio y video, por ejemplo o gran tráfico de datos. También se recomienda el uso de cables de categoría 6 (Gigabit Ethernet) que permiten el aumento de la velocidad de la red LAN hasta 250 Mbit/s, aumentando considerablemente su velocidad.

Con las mejoras en la red, si fuera necesario se podría aumentar la frecuencia de muestreo y la codificación de las señales de audio. No obstante, los valores calculados son más que suficientes para asegurar una representación correcta de las señales de voz que permien a la aplicación de reconocimiento un funcionamiento correcto.

4.2.3 Uso de una red inalámbrica para la transmisión de las señales

Dado el cada vez más extendido uso de las redes inalámbricas es necesario plantear la posibilidad de utilizar una red inalámbrica en lugar de una red cableada. Desde el punto de vista técnico la modificación respecto al sistema de LAN cableada es la desaparición del cable ya que el medio portador es ahora el aire. Para ello, el adaptador será del tipo inalámbrico WLAN.

Este tipo de red tiene ventajas respecto al cableado en el caso de instalaciones de viviendas ya construidas y en las que no es posible realizar obras de adecuación para la instalación de canalizaciones y registros. La desventaja respecto al sistema cableado es su fiabilidad. Un sistema vía radio puede verse influido por el entorno y su fiabilidad disminuye si se aumenta la distancia a cubrir. Es necesario valorar adecuadamente su utilización a fin de evitar problemas de funcionamiento derivados de falta de conectividad temporal por otras redes cercanas, de similares características, o por inhibidores de frecuencias utilizados por las fuerzas de seguridad para la protección de personas públicas.

Hay que distinguir entre una red inalámbrica que utiliza captadores instalados de forma fija en cada estancia (micrófonos situados en techo o paredes) de una red inalámbrica que utiliza un dispositivo portátil, del estilo de un mando a distancia que puede llevar el usuario consigo. Este dispositivo puede ser un emisor-receptor del estilo de un walkie-talkie, o un mando a distancia o incluso una PDA. Este sistema se define como conexión inalámbrica móvil.

De los sistemas descritos anteriormente pueden derivarse combinaciones en la instalación, por ejemplo: sistema de captadores fijos conectados a red LAN dentro de una vivienda unifamiliar en la que se usa un emisor-receptor portátil cuando se está en el jardín de la vivienda que se conecta a la red mediante un punto de acceso inalámbrico.

4.2.4 Integración con sistemas domóticos comerciales

Se presentan a continuación las conclusiones del análisis de varios sistemas comerciales existentes de control domótico respecto a la posibilidad de utilizar los BUSES de dichos sistemas para el transporte de las señales de audio del sistema de reconocimiento y síntesis de voz. Según el punto anterior, una tasa binaria de referencia para que la transmisión de los datos del sistema funcione correctamente, debe ser al menos de 1,6 Mbit/s.

4.2.4.1 KNX (EIB) y LONWORKS

- En ambos casos se permite la posibilidad de transmitir información (no propia del protocolo) entre dispositivos conectados a dicho BUS (textos en pantallas, temperaturas, etc.).

- La velocidad del BUS en ambos casos (9600 bit/s en KNX y 7800 bit/s en LonWorks), resulta insuficiente para las necesidades de transmisión mínimas que requiere el sistema de reconocimiento y respuesta por voz.

4.2.4.2 Sistemas basados en corrientes portadoras (X10)

- La tasa binaria de estos sistemas está relacionada con la frecuencia de la red eléctrica (50Hz o 60Hz). Esto, unido a redundancias necesarias para asegurar la correcta recepción de los mensajes hace que la tasa binaria sea muy baja y no sean sistemas preparados para la transmisión de grandes volúmenes de datos. Por tanto no es posible utilizar estos protocolos para la transmisión de los datos de la aplicación de control por voz.

4.2.4.3 ZigBEE

- La propia filosofía del sistema inalámbrico, en la que el consumo y la transmisión de datos (250 Kbit/s) deben minimizarse, hace inviable la utilización del protocolo propio para el transporte de las señales de audio hacia la unidad central de procesamiento.

4.2.4.4 Z-Wave

- Este sistema inalámbrico puede funcionar a dos velocidades binarias, 9600 bit/s o a 40 Kbit/s, ambas insuficientes para los requerimientos del sistema de voz.

4.2.4.5 BUSing

El sistema está desarrollado para su funcionamiento sobre varios medios de transmisión: RS-485, CAN, TCP/IP y radio (868 MHz y 2,4 GHz).

- Según la especificación del RS-485, la velocidad binaria que se puede conseguir depende inversamente de la longitud del cable que se utiliza. Para distancias cortas (menores de 50 m) es posible disponer de tasas binarias por encima de 18Mbit/s. Esta tasa binaria en principio está por encima de los requerimientos para el transporte de las señales de audio. Hay que tener en cuenta la tasa binaria propia del protocolo BUSing. Esta tasa binaria dependerá de los dispositivos domóticos conectados y su actividad. En general es conveniente evitar la carga excesiva de datos en el bus y la ralentización de la respuesta del sistema.

Queda como posible estudio futuro la forma óptima de encapsular y compatibilizar el sistema de control por voz con el sistema BUSing funcionando sobre RS485 y sus limitaciones de utilización.

- El Bus CAN permite velocidades de datos de hasta 1 Mbit/s por lo que sería insuficiente para el transporte del las señales del protocolo BUSing y de las señales del sistema de control por voz.

- En el caso de redes TCP/IP ya se ha comprobado la viabilidad para el transporte de múltiples datos en una red que funciona como mínimo a 100 Mbit/s.

- En el caso de las redes vía radio las velocidades binarias efectivas bajan considerablemente debido principalmente a los algoritmos de corrección de errores necesarios por las interferencias inherentes al canal de radio. En este caso, al igual que ocurría con Zigbee y Z-wave la tasa binaria no es suficiente.

4.2.4.6 Powerline-Ethernet. PLC

- Las tasas binarias que consiguen los dispositivos que extienden la red LAN mediante el uso del cableado eléctrico (hasta 200 Mbit/s) permiten la utilización de la red eléctrica para el transporte de las señales. Un fabricante de este tipo de dispositivos adaptadores es COMTREND.

CONCLUSIONES:

- La mayoría de los sistemas domóticos están diseñados para utilizar tasas binarias bajas y por tanto no es posible encapsular volúmenes de datos grandes en su protocolo de comunicaciones.

- Siempre que exista una red LAN, cableada, inalámbrica o transmitida mediante adaptadores PLC por el cableado eléctrico, es conveniente llevar las señales de audio procedentes de los captadores de audio hasta la unidad central.

RECOMENDACIÓNES:

1. Es conveniente que el sistema de control por voz transporte sus señales de audio de forma independiente al del sistema domótico. Debe disponer de un interfaz adicional (Adaptador a bus) que permita la interconexión con el sistema domótico (cableado o inalámbrico) en ambos sentidos (bidireccional).

2. En vivienda ya construida en la que resulta difícil introducir cableado nuevo, se pueden utilizar adaptadores de línea eléctrica-Ethernet (PLC) para la extensión de la red LAN.

4.3 Ventajas e inconvenientes según el tipo de conexión

El punto más crítico que condiciona el funcionamiento del sistema es la captación de las señales de audio procedentes del hablante. Los elementos captadores son sistemas compuestos por micrófonos y previos que pre-amplifican la señal para su correcta propagación hasta el sistema central de procesado. La conexión de los captadores se puede realizar de dos formas, cableada e inalámbrica.

TIPO DE CONEXIÓN DE CAPTADORES	
CABLEADA	EN ESTRELLA
	POR BUS
INALÁMBRICA	FIJA
	MÓVIL

TABLA 1

Una conexión cableada en estrella conectará cada uno de los elementos captadores mediante un cable dedicado hasta la entrada correspondiente del sistema de reconocimiento.

En una conexión cableada por BUS el sistema de captación debe disponer de la electrónica y el firmware necesario para implementar los protocolos de comunicación (prioridad de acceso, velocidad, errores) para la transmisión de las señales captadas a través del BUS en el que se encuentra conectada la unidad de procesamiento. Este BUS puede ser el mismo destinado al control del Hogar Digital siempre que cumpla con los requisitos de transmisión de la aplicación de voz.

Cuando se utilizan captadores inalámbricos, la transmisión desde el captador se realiza por RF. Si el captador está instalado en la vivienda se considera fijo. Si el captador es un dispositivo que el usuario puede llevar consigo se considera móvil. En el caso del móvil inalámbrico se pueden además considerar dos tipos, el primero, aquel que únicamente

transmite y recibe audio por medio de RF, y el segundo que incorpora el reconocimiento y síntesis en el propio dispositivo.

Este segundo dispositivo con el sistema de reconocimiento y síntesis de voz incorporado, debe disponer de una conexión inalámbrica para su interconexión con los elementos que controla. Otra gran desventaja de esta tipología es el alto consumo de batería del dispositivo móvil. Si se aumenta el tamaño de la batería su peso y dimensiones pueden aumentar considerablemente con la pérdida de facilidad en la portabilidad.

Se resume en la tabla siguiente una comparativa de las tipologías genéricas de conexión entre captadores y sistema central.

TIPO DE CONEXIÓN	VENTAJAS	INCONVENIENTES	USO
FIJA CABLEADA	Seguridad y Robustez	Reconocimiento dependiente de la ubicación de los captadores	Nueva Construcción
FIJA INALÁMBRICA	Facilidad en la instalación	Posibles interferencias. Menos robusto	Vivienda construida
MÓVIL	Reconocimiento óptimo	Consumo. Necesidad de recarga	Vivienda construida.

TABLA 2

4.4 Alimentación de los sistemas y consumo.

La alimentación de los sistemas que se instalan dentro de la vivienda es un tema que hay que considerar desde la fase de diseño. En el caso del sistema de control por voz cableado mediante bus existe la posibilidad de llevar la alimentación de los dispositivos adaptadores de bus por el propio cable de BUS. Este es el caso de las redes Ethernet llamadas PoE (Power Over Ethernet o Power over LAN). Para ello es necesario que los elementos de enrutamiento dispongan de esta tecnología. Otro tipo de sistemas de Hogar Digital en los que se puede utilizar la alimentación que proporciona el BUS es el estándar europeo KNX. En cualquier caso es necesario comprobar que los sistemas adicionales que se conectan al BUS no demanden más corriente que la permitida por el BUS.

En el resto de casos en el que el propio cable del BUS no pueda suministrar la alimentación a los dispositivos conectados a el, es necesario suministrar aparte la alimentación a los dispositivos captadores y emisores. Para estos casos se debe disponer de una fuente de alimentación AC/DC situada en el armario donde estará instalada la unidad central. Desde el punto donde está situada la fuente hasta cada dispositivo a alimentar, será necesario llevar un cable de alimentación. La fuente de alimentación, en todos los casos debe dimensionarse para que el consumo de los elementos captadores y emisores sea menor que el máximo que la fuente puede suministrar con un margen de seguridad amplio. Hay que tener en cuenta que los altavoces pueden tener consumo elevado en comparación con dispositivos electrónicos de red LAN. En una vivienda en la

que por lo general no es necesario disponer de gran potencia sonora unos altavoces tipo multimedia con potencia 10W pueden ser más que suficiente para que el nivel de audio esté en unos valores adecuados.

CONSIDERACIÓN:

- Cabe destacar en este punto la posibilidad de utilizar el cableado eléctrico para la transmisión de señales mediante el uso de PLC. Esta técnica aporta una ventaja importante: la eliminación de cableado adicional, tanto para la transmisión de señales, como para la alimentación de los equipos electrónicos. En toda vivienda, nueva o ya construida, se puede aprovechar la instalación del cableado eléctrico que discurre por el techo de cada estancia para incorporar un pequeño adaptador (que podría quedar oculto en la mayoría de los casos por la lámpara) cuya finalidad sería, por un lado, dar alimentación a una placa captadora (micrófono más previo) situada en el techo, y por otro, actuar de transductor-adaptador LAN de las señales captadas por el micrófono. La única consideración técnica a tener en cuenta con esta técnica, es la necesidad de incluir un cable adicional que lleve la fase eléctrica desde antes del interruptor que enciende o apaga la luz, hasta la salida de cables que conecta la propia luz. De esta forma se asegura que la alimentación del dispositivo captador es independiente del estado del interruptor. Igualmente podría estudiarse la misma solución para los elementos emisores.

4.5 Interfaz con sistemas de Hogar Digital

El sistema de control por voz requiere de un procesado de señal potente en relación con los procesos de actuación y regulación que en la mayoría de los casos son necesarios para el control domótico de automatismos de la vivienda. Por este motivo el sistema de procesado, que hemos llamado unidad central debe ser un hardware específico: un sistema DSP dedicado o una plataforma hardware con software específico que implemente el reconocimiento y la síntesis de voz. Un módulo adicional debe actuar de interfaz entre el sistema de control por voz y el sistema domótico al que se asocia.

En el caso de KNX, LONWORKS o X10 se puede desarrollar un interfaz o pasarela compatible con los protocolos de comunicaciones utilizados por estos sistemas o bien se puede hacer uso de un interfaz estándar comercial que transforme datos serie (RS485, RS232) al BUS del sistema de control domótico. Para sistemas inalámbricos Zigbee, Z-Wave, el interfaz de unión con el sistema de voz será inalámbrico.

En el caso del sistema UPNP que funciona sobre redes LAN o WAN el propio protocolo tiene definida la forma de encapsular o empaquetar el audio permitiendo la realización de streaming a tiempo real. En estos casos el sistema dispondrá de una tarjeta de red Ethernet.

En términos comparativos con ordenadores personales, se comprueba que el hardware mínimo para realizar reconocimiento y síntesis de voz puede ser equivalente a un microprocesador Pentium IV con una memoria RAM de 1 GB. Actualmente cualquier ordenador personal supera ampliamente esas características técnicas. Existen soluciones

de ordenadores industriales con versiones reducidas de sistemas operativos (Windows XP Embedded) idóneas para implementar estas aplicaciones.

4.6 Ruido ambiente y reconocimiento robusto

El ruido ambiente es el mayor enemigo del reconocimiento de voz. Para aplicaciones en las que el dispositivo de captura está cerca del locutor o bien el ruido ambiente es bajo no hay demasiados problemas puesto que el nivel de voz (señal a reconocer) respecto al ruido es alto (relación señal-ruido alta). En casos con relaciones señal ruido bajas se hace necesario el uso de técnicas de reducción de ruido. Tradicionalmente se han usado técnicas como filtros de Wiener, o de menor carga computacional y menos "fina" como la sustracción espectral.

Además del problema que surge al tener el ruido ambiente o señales no deseadas, mezclados con la señal a reconocer, aparece otro efecto: la persona que se encuentra en un ambiente ruidoso cambia su forma de hablar forzando la voz para poder escucharse mejor. Al forzar la voz las características de esta varían sustancialmente. Este efecto debe tenerse en cuenta en el reconocedor. Para ello el corpus de voz a utilizar debe disponer de muestras de voz en condiciones de ruido ambiente alto con varios niveles de intensidad sonora.

Otro efecto a tener en cuenta es el de la reverberación de los recintos. Este efecto hace que aparezcan múltiples señales similares a la original que retardadas y filtradas aparecen mezcladas con la señal original. Este efecto es muy evidente en habitaciones vacías en las que existe ningún elemento que amortigüe las reflexiones.

Al tratarse de captura de señales de voz, los sistemas de captación pueden filtrar señales de audio que estén fuera del rango de la voz humana. De esta forma, señales de baja frecuencia, por debajo de 100 Hz, y señales de frecuencia alta, por encima de 8.000Hz, que no aportan información del habla pueden directamente ser eliminadas con un filtro paso banda en el propio elemento captador.

En el hogar pueden existir múltiples fuentes de ruido que afectarán negativamente a la captura:

- Electrodomésticos: lavadora, secadora, lavavajillas.
- Sistemas de audio-video: televisor, Equipo HI-FI, radios y reproductores de audio.
- Ventanas abiertas a la calle.
- Conversaciones de varias personas diferentes a la que se quiere reconocer.
- Mascotas y animales de compañía: perros, gatos, pájaros.
- Conversaciones de otras personas en la misma estancia.

Para minimizar al máximo los efectos negativos de las fuentes de ruido indicadas se analiza a continuación las recomendaciones de funcionamiento e instalación.

4.6.1 Instalación de los elementos captadores

Una buena ubicación de los elementos captadores en función del uso al que se destina la estancia mejora notablemente el proceso de captura. El caso peor se da si el sistema se va a diseñar sobre planos de vivienda nueva. Es necesario conocer previamente la disposición de parte del mobiliario de la vivienda. Por ejemplo, a la hora de instalar el captador en un dormitorio principal es necesario conocer la disposición del cabecero de la cama. En la fase de diseño arquitectónico se define la disposición del mobiliario. En función de esta disposición los ingenieros de telecomunicaciones e ingenieros eléctricos indican en sus proyectos las tomas de telecomunicaciones y eléctricas a ambos lados de la cama. El mismo caso se da con las tomas de telecomunicaciones en los salones y comedores. El arquitecto indica el lugar donde se instalará la televisión, y en ese punto se disponen tomas de telecomunicaciones y de electricidad. Si se ha establecido la zona para ver la televisión y el lugar donde se instalará la televisión, es seguro que los usuarios siempre se van a sentar mirando hacia la televisión. Por tanto se define una zona de mayor probabilidad donde instalar los captadores.

Teniendo en cuenta las consideraciones anteriores, y en coordinación con los técnicos que diseñan las infraestructuras dentro de cada vivienda, se proyectarán las canalizaciones y registros necesarios para el control por voz.

Hay que evitar colocar los captadores cerca de fuentes de ruido, por ejemplo, en una cocina no conviene colocar el captador cerca de la campana de extracción de humos. Igualmente, no convendrá instalar los elementos captadores cerca de los altavoces de televisores o de equipos de audio.

Existen soluciones comerciales formadas por arrays de micrófonos con forma de aplique con un deflector que cuelga del techo, diseñados para captar de forma omnidireccional las señales de voz, ajustando ganancias y evitando ecos. Están pensados para aplicaciones de manos libres: conferencias, salas de reuniones, etc. Son soluciones caras y estéticamente no son muy adecuados para instalar en todas las estancias de en una vivienda.

Como solución para la mejora del ruido de ambiente puede utilizarse en cada estancia un micrófono adicional asociado al mismo elemento electrónico captador. La función de este segundo micrófono, alejado del primero, es la de recoger el ruido ambiente o señales interferentes y realizar una mejora en la calidad de la señal que lleva la información de voz

La forma clásica de reducción de ruido ambiente en sistemas portátiles en ambientes ruidosos ha sido básicamente la sustracción de la señal procedente del micrófono secundario si este sólo registraba la señal interferente. En la práctica no se consigue separar completamente las señales de ruido de la señal de voz. En el dominio digital, básicamente la reducción del ruido ambiente puede realizarse mediante la implementación de un filtro en celosía que detecta las características espectrales básicas de la señal de ruido y a continuación un filtro en escalera que realiza una sustracción de la señal de ruido de la señal a reconocer de forma ponderada. En estos casos se puede mejorar la relación señal ruido en más de 10 dB.

Como mejora, y sólo a efectos de conocer el estado del arte de estas técnicas de la eliminación de ruido se está experimentando con la llamada Supresión de Ruido Audible (Audible Noise Supresión). Esta técnica tiene que ver con la forma en que el sistema auditivo humano interpreta el ruido y cómo se enmascaran los sonidos. El oído humano tiene unas bandas de audición, llamadas bandas críticas. Utilizando los umbrales de enmascaramiento, definidos para cada una de las bandas críticas, se puede conocer cuál es la cota superior, por debajo de la cual el ruido no afecta a la señal. De esta forma, si en una banda crítica, ese umbral de enmascaramiento es alto, no es necesaria una supresión de ruido alta y viceversa. Aplicando la reducción selectiva de ruido de esta forma, la respuesta del reconocedor mejora notablemente y la tasa de reconocimiento aumenta en presencia de ruido.

En el caso de la tipología de BUS en el que se instala una unidad de adaptación a BUS (ADB) que dispone de un microcontrolador, es posible realizar parte de esta mejora de la señal: supresión de ecos y reducción de ruido como parte de preprocesado de audio y de esta forma descargar a la unidad central de estos cálculos.

Si las fuentes de ruido son conocidas, por ejemplo, el sonido del televisor o del equipo de música, y es posible llevar la señal de audio hasta el sistema de procesado se mejora notablemente la recepción puesto que el sistema será capaz de reconocer dicha señal como interferente o señal no válida. Esto se puede conseguir situando un micrófono cerca de las fuentes de ruido conocidas. De esta forma la estimación de la señal de ruido será mucho más efectiva.

Habrá ciertas condiciones, como la comentada en el párrafo anterior en las que el sistema de reconocimiento puede no entender lo que se le dice. Si el usuario entiende que el sistema falla al igual que ocurre con los humanos ante una presencia de un ruido muy alto o ante varias conversaciones a la vez o ante un televisor con excesivo volumen, será capaz de solucionar el problema de la misma forma que lo haría si él mismo quisiera entender lo que se le dice: cerrando la ventana, mandando callar o bajando el volumen.

Una situación que puede producirse con frecuencia es la presencia de personas conversando en la misma habitación en la que se está realizando una comunicación con el sistema. En estos casos, dependiendo del volumen de la conversación y la distancia al sistema captador es posible que el reconocimiento falle. Al igual que ocurre en nuestras comunicaciones, será necesario que cese la conversación para poder continuar con la comunicación.

CONCLUSIONES:

- Una elección acertada de la instalación de los micrófonos en las estancias mejora notablemente el funcionamiento del sistema.

- Existen zonas de mayor probabilidad de reconocimiento del captador relacionadas con la ubicación del mobiliario de cada estancia y su utilidad. Esto condiciona la instalación de los captadores.

- Ciertas situaciones son no deseables para un reconocimiento correcto. Es necesario evitar fuentes de ruido de alto nivel (ruido de tráfico rodado, equipos de HIFI o televisores o conversaciones de personas) si se quiere realizar una comunicación cómoda y fiable con el sistema.

4.7 Palabra de atención. Humanización del sistema

En nuestra vida cotidiana la forma de requerir la atención de alguien suele ser llamarle por su propio nombre. Es la forma más común de comenzar una petición o dar una orden. Por tanto lo más lógico es que nuestro sistema de control por voz del Hogar Digital se adapte a nuestra formas de comunicación y no al contrario. Seguro que al lector le vienen a la cabeza películas de ciencia ficción como: 2.001: Odisea del espacio (1968) donde HAL 9000 era el ordenador de abordo con el que se comunicaban los tripulantes de la nave Discovery mediante voz y con el que mantenían conversaciones de toda índole. Otra célebre película de ciencia ficción es Alien, el octavo pasajero (1979) donde la computadora central se llama "Madre" y con la que conversan para dar órdenes a la nave espacial Nostromo.

Aunque parece trivial, la elección de una palabra de atención para comenzar la comunicación con el sistema debe ser pensada y elegida por el usuario cuidadosamente. No deben permitirse palabras comunes usadas normalmente en conversaciones cotidianas dentro del hogar. Es recomendable el uso de nombres poco comunes y de más de dos sílabas. Unos ejemplos pueden ser: Mayordomo, Ambrosio, Recaredo, Romualdo, etc…

En función de la palabra de atención y su género el usuario puede elegir entre varias voces sintéticas de respuesta diferentes, tanto masculinas como femeninas.

El hecho de que el sistema reconozca una palabra de atención permite que la aplicación permanezca en stand-by hasta ser llamada. Otra ventaja respecto al procesado es que el sistema puede hacer un ajuste de ganancia y eco con la palabra de atención (comparándola con la almacenada en su base de datos en condiciones normales) previamente al reconocimiento de las órdenes posteriores.

Siguiendo la forma de comunicación verbal, las personas, tras entender nuestro nombre solemos contestar con: ¿sí, dime?, ¿qué deseas?, ¿en qué puedo ayudarte?, etc. El sistema imita este comportamiento para que el usuario sepa que su llamada de atención ha sido entendida. La elección de la frase de respuesta debe poderse modificar en función de las preferencias de cada usuario.

Un punto clave es el tiempo de respuesta, es decir, el que pasa desde que se ha terminado la palabra de atención hasta que el sistema contesta mediante síntesis de voz o mediante la activación de algún interfaz de salida. Se ha comprobado mediante diferentes estudios de interfaces y de sistemas de control, que tiempos superiores a 1 segundo en la respuesta no son aceptables, e implican que el usuario vuelva a formular la petición de atención, ya con cierta frustración por la tardanza en la respuesta. Este parámetro es

crítico en el desarrollo del sistema y viene condicionado por la potencia de procesado y por la integración óptima con el sistema domótico. Igualmente hay que tener en cuenta que una vez establecida la comunicación, si el usuario da una orden, por ejemplo, de encendido de una luz, el tiempo de actuación no debe superar el segundo para no transmitir al usuario sensación de fallo o lentitud del sistema. Por tanto es recomendable trabajar con tiempos de respuesta bajos. En el caso de sistemas de BUS deben evaluarse previamente todos los tiempos de retardo para trazar correctamente el BUS y su conexionado.

Un complemento a la síntesis de voz para una mayor humanización de los sistemas de síntesis de voz combinado con imágenes es la utilización de un personaje virtual o también llamado avatar. Se muestran a continuación 3 ejemplos de avatares desarrollados por la empresa Indisys.

FIGURA 9

Para ello se hace uso de los elementos visuales disponibles asociados al sistema: pantallas táctiles, televisores, ordenadores y PDAs. En dichas pantallas, donde se encuentra el interfaz gráfico de control del sistema, se genera un personaje gráfico mediante técnicas de modelado 3D haciendo que aparezca una figura que con la que se expresa gestualmente con el usuario como si se tratara de una conversación entre dos personas. Una implementación conductual de estos personajes que se realiza es que cuando el sistema está inactivo, el avatar aparece dormido. Cuando se llama al sistema por su palabra de control el avatar se despierta y dirige su mirada hacia el usuario. Esta forma de comportamiento transmite sensación de privacidad al usuario, de tal forma que el avatar no escucha si no es llamado.

En sistemas de gestión de lenguaje natural estos avatares ganan protagonismo ya que transmiten y reciben emociones de una forma muy real, humanizando la interacción con el usuario.

En los casos en que el usuario esté hablando con el sistema y se produzca un ruido espontáneo de alto nivel que afecte a significativamente al reconocimiento (el timbre, un teléfono, una sirena, etc.) el sistema puede indicar al usuario que repita la orden. Normalmente en el transcurso de una conversación cuando esto ocurre es lo que las personas hacemos normalmente.

CONCLUSIÓN:

- Cuanto más se humanice el sistema menor rechazo existirá por parte del usuario consiguiendo una eficacia mayor en el uso.

4.8 Verificación del hablante. Identificación biométrica

Hasta hace poco tiempo la verificación del hablante mediante huella de voz no era posible. Sistemas de reconocimiento de voz comerciales como Loquendo ya permiten su realización por lo que no es difícil aventurar que su utilización como medida biométrica se extenderá en breve al uso de sistemas de control por voz. Las aplicaciones en el Hogar Digital son variadas. Una aplicación con claros beneficios es el establecimiento de prioridades dentro del Hogar, en el que determinados usuarios dentro del hogar pueden tener un control completo mientras que otros pueden restricciones. En el momento que el sistema reconoce la huella de voz del hablante, comprueba sus permisos, previamente asignados y actúa en consecuencia, o bien accediendo a la petición, o bien notificando la denegación de la petición por falta de permisos.

Haciendo uso de esta tecnología de huella de voz cada usuario podrá interactuar con el sistema de forma diferente. El sistema, al identificar al hablante, actuará en función del perfil de dicho usuario: palabra de atención, frase de respuesta, timbre de voz, preferencias, etc. Por ejemplo, el sistema al entrar un usuario a la vivienda le dará la bienvenida realizando alguna pregunta de control. Cuando el usuario conteste, el sistema, al reconocer al hablante podrá actuar sobre la iluminación, o el equipo de música, según las preferencias que cada usuario había fijado previamente como escena de "llegada a casa".

4.9 Localización de los usuarios

En todo momento el sistema debe saber en qué estancia se encuentra el usuario y debe tener cierta información sobre los dispositivos a controlar en dicha estancia. Por ejemplo:

El usuario llama a: "*Ambrosio*", y espera respuesta

"*Díme que puedo hacer por ti, Fernando*", responde el sistema al usuario habiendo analizado la voz del hablante y contestado según los gustos de ese usuario del hogar.

"*Baja la persiana y enciende la luz.*" Indica el usuario.

En este punto el sistema conoce en qué estancia está el usuario ya que el audio se ha encaminado desde el sistema captador de la estancia salón,

¿Qué persiana bajo? .El sistema sabe que hay dos persianas en el salón y pregunta para completar el resto de información necesaria antes de actuar.

"La persiana de la terraza". El usuario ha nombrado esa persiana de la forma que el sistema conoce.

El sistema puede, en función de una programación previa del usuario, preguntar cuánto quiere bajar la persiana, o simplemente, ante la orden, "bajar la persiana" interpretar que debe bajarse hasta el final de carrera. A continuación, *Ambrosio* comprueba qué luces se encuentran encendidas en el salón y, en el caso de que hubiera varias, de igual forma que ha preguntado para las persianas, solicitaría más información. Si sólo hay una luz la encenderá.

Si además el usuario quiere realizar una acción en una estancia en la que no se encuentra deberá nombrar dicha estancia: *"Ambrosio, enciende la luz del recibidor".*

"luz del recibidor encendida". Contestará el sistema, dando una confirmación de la acción realizada.

Una utilidad de esta aplicación es la posibilidad de que el sistema retransmita mensajes de los usuarios entre estancias, por ejemplo:

"Ambrosio, di los niños que ya está la comida".

El sistema ha recordado cuáles han sido las últimas estancias donde los niños han estado hablando con el y en esas estancias emite el mensaje indicando el mensaje de la madre.

También es posible localizar a usuarios:

"Ambrosio, localiza a Fernando"

El sistema emitirá un mensaje de broadcast por toda la casa indicando al usuario que le están buscando. En el momento que el usuario buscado conteste, el sistema localiza su ubicación, e indica al usuario que ha generado la consulta, la estancia en que se encuentra la persona buscada. Como continuación a esta situación los usuarios podrían establecer una comunicación por voz entre ambas estancias como si se tratara de un intercomunicador manos libres simplemente ordenando al sistema entrar en modo intercomunicador.

4.10 Tareas en paralelo

A priori la limitación para realizar tareas en paralelo respecto al reconocimiento y síntesis de voz recae en la capacidad de proceso del hardware que ejecuta la aplicación. En principio no parece una limitación importante ya que los sistemas hardware actuales son cada vez más potentes: más memoria RAM, y procesadores dobles. Por tanto, un sistema de control por voz que se ejecute sobre un hardware similar será capaz de atender

a dos o más usuarios, cada uno en una estancia de la vivienda hablando simultáneamente. Existirán condiciones de funcionamiento a tener en cuenta a la hora de procesar órdenes simultáneas que modifican el estado de un mismo elemento de la vivienda. Por ejemplo: dos usuarios pueden encontrarse en habitaciones diferentes y dar una orden de actuación sobre una luz o una persiana de forma simultánea. En ese caso el sistema debe notificar a uno de ellos que ya hay otro usuario ordenando un cambio de estado sobre ese elemento.

5 SISTEMAS Y PRODUCTOS COMERCIALES

Toda la información expuesta en el punto anterior ha sido extraída de la investigación realizada de los sistemas actuales de control por voz. Si bien ninguno reúne a día de hoy todas las características descritas en los apartados anteriores, cada uno aporta su contribución al estado del arte de esta tecnología, yendo un paso más allá y poniéndola en manos de los usuarios. Se describen a continuación algunos de los sistemas examinados intentando centrarnos en sus ventajas, inconvenientes y justificando el campo de aplicación para el que han sido diseñados. En la mayoría de los casos son productos de reciente aparición, o incluso están en fase de desarrollo y mejora.

5.1 Fagor. Maior Vocce

Desde hace años Fagor dispone de un sistema de control del Hogar Digital denominado Maior Domo. Este sistema se basa en el uso de corrientes portadoras que aprovechan el cableado eléctrico de la vivienda para transportar las señales de control. El sistema permite el control de automatismos, electrodomésticos, la gestión energética y la seguridad, tanto técnica como anti-intrusión.

FIGURA 10

El control por voz se añade para permitir controlar estos dispositivos mediante la voz. El sistema de control por voz consiste en un dispositivo portátil que el usuario puede llevarlo a modo de reloj o en forma de colgante. Se comunica vía radio con un emisor-receptor inalámbrico que se instala en carril DIN (cuadro eléctrico) que se comunica con el resto de la instalación mediante los códigos de control que se transmitirán por la red eléctrica. Además dispone de un interfaz serie RS485 para conexionado con otros dispositivos.

La pulsera o colgante transmite y recibe las señales a una frecuencia de 868 MHz con una potencia radiada máxima de 25mW. Utiliza una batería recargable de Ion-Litio de 3,7V y una capacidad de 180mA/h.

El sistema de reconocimiento de voz está basado en menú de comandos y es independiente del hablante, es decir, reconoce las órdenes de cualquier tipo de persona sin entrenamiento previo.

El tipo de usuario al que se orienta es el de colectivos de la tercera edad y personas dependientes. El precio del sistema que incluye la pulsera-colgante RF y el receptor-emisor RF que interconecta con el sistema domótico es de 800 €.

Ventajas:

- Sistema sencillo de instalación y mantenimiento.
- Robusto al ruido ambiente: captación cercana no hay problemas de fallo en reconocimiento y menú de comandos fijo con palabras predefinidas.
- Integrado con sistema de Hogar Digital propio, eliminando incompatibilidades.

Inconvenientes:

- Necesidad de aprendizaje por parte del usuario de los comandos de voz a utilizar. Problemas de memoria en personas de la tercera edad.
- Necesidad de recarga de batería.
- Comunicación vía radio, sujeta a interferencias y posibles fallos en la comunicación.

5.2 Proinssa

La empresa orienta el control por voz hacia personas dependientes o personas de la tercera edad. Disponen de interfaces y aplicaciones específicas para discapacitados en entornos hospitalarios y de hogar accesible. Sus interfaces para el control por voz se basan en dispositivos portátiles del tipo mando a distancia programable que es capaz de copiar los códigos de infrarrojos de otros mandos a distancia. Pueden accionarse mediante pulsación o mediante la voz. El sistema de control por voz se integra en el mando a distancia. Es del tipo reconocedor de comandos. Las opciones de voz se estructuran por menús. Aunque el sistema reconoce comandos independientemente del usuario, para una mejora en el reconocimiento o para usuarios con algún problema de fonación, es conveniente grabar previamente varias muestras de voz para cada comando del menú. Cuando el mando a distancia reconoce la palabra asociada al comando envía mediante infrarrojos la orden correspondiente a los receptores que controlan los sistemas a accionar: motores de puertas, camas, luces o persianas. Disponen de dispositivos propios con receptor de infrarrojos y actuador para motores de camas, enchufes eléctricos, motores de puertas y ventanas, persianas y pasa páginas.

También permiten, mediante el uso del interfaz adecuado la integración con otros sistemas domóticos comerciales: KNX, LONWORKS o X10.

FIGURA 11

En los casos de personas minusválidas, la instalación puede realizarse en la cama o en la silla de ruedas, dejando el receptor como un elemento fijo.

El mando cuenta con una base para la recarga. Se recomienda cargarlo cada noche, si bien el mando puede funcionar sin problemas de batería por varios días.

Permite el ajuste de ganancia en función de la distancia y del volumen de la voz a reconocer. Para entornos ruidosos o personas con problemas de fonación permite el uso de un micrófono de tipo diadema.

Disponen de 3 tipos de mando SICARE: LIGHT, BASIC y STANDARD. En el primer caso el LIGHT el menú de comandos es fijo, en el segundo caso, el BASIC es posible cambiar mediante programación vía ordenador, el menú de comandos, tanto el orden como el dispositivo. En el último caso se cuenta con una posibilidad de transmisión vía RF que debe asociarse a dispositivos propios receptores-actuadores. Los precios de los tres mandos varían sustancialmente, el SICARE LIGHT tiene un PVP de 1600 €, el SICARE BASIC, 3500 € y el SICARE STANDARD 5.000 €.

Ventajas:

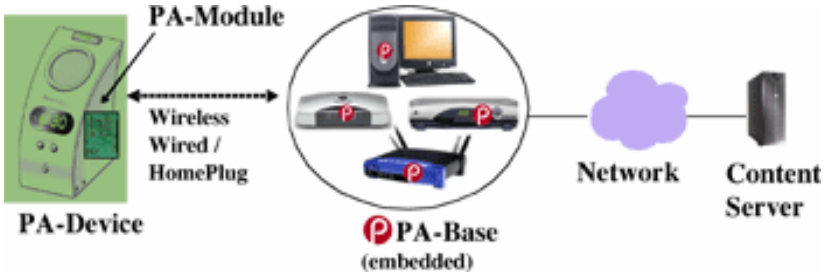
- Sistema sencillo de instalación y mantenimiento.
- Robusto al ruido ambiente: captación cercana (mayor efectividad de reconocimiento) y menú de comandos fijo con palabras predefinidas.
- Permite integración con otros sistemas: KNX, LONWORKS, X10

Inconvenientes:

- Necesidad de aprendizaje por parte del usuario de los comandos de voz a utilizar. (Problemas de memoria en personas de la tercera edad).
- Necesidad de recarga de batería.
- Transmisión vía infrarrojos que no es omnidireccional: necesita apuntar hacia el receptor (posibles fallos en la recepción).

5.3 Personica

Personica es una empresa estadounidense, situada en el área de Boston pionera en el desarrollo de tecnología de "Asistencia Personalizada". Abogan por el uso fácil de las tecnologías de Hogar Digital para que su implantación sea definitiva. Para ello hacen uso de tecnologías de reconocimiento de voz para el control de sistemas basados en redes LAN.



Basan el reconocimiento de voz en unas tarjetas llamadas PAM (PA-Module). Estas tarjetas implementan el procesado para la captación, con posibilidad de conexión de varios micrófonos para mejora de la señal. Además permiten la comunicación con el bus para recibir la síntesis de voz del sistema de control. La tarjeta que realiza al completo las funciones de reconocimiento de voz, permitiendo la identificación biométrica (reconocimiento de hablante). Su activación se realiza con una palabra de atención a una distancia máxima de 10m en presencia de ruido ambiente. Las tarjetas se comercializan para su integración en dispositivos finales. Dependiendo del acabado final se pueden incluir amplificadores y altavoces al dispositivo definitivo. De esta forma se tiene un único dispositivo que realiza tanto la captación como la emisión de voz.

La otra parte del sistema se realiza mediante una aplicación software, PA-Base Framework, que implementa la síntesis y el análisis de voz y la interconexión con el sistema domótico. Está pensado para la instalación en Home Media PCs. Este software está disponible para su instalación bajo sistemas Windows o sistemas Linux. Este software permite la instalación de aplicaciones adicionales que mejoran el sistema: librerías de lenguaje preferencial o aplicaciones de reconocimiento de voz en lenguaje natural basado en el contexto.

El tipo de mercado al que se orienta es aquel de viviendas de alta gama, viviendas unifamiliares con instalaciones de telecomunicaciones, audio y video integradas utilizando redes LAN. El interfaz soporta la tecnología UPNP.

Ventajas:

- Comunicación aprovechando redes LAN, cableadas o inalámbricas.
- Manos libres y robustas al ruido (10 m de alcance en reconocimiento con ruido)
- Identificación del hablante (huella de voz)
- Reconocimiento y síntesis de lenguaje natural con interpretación según contexto.

Inconvenientes:

- Actualmente no es un sistema acabado, listo para instalación, sino que debe ser integrado con sistemas de Hogar Digital.
- El que cada tarjeta (que debe instalarse en cada estancia) incorpore todo el procesado de audio hace muy probable que su precio sea alto, pudiendo ser un lastre en la comercialización del sistema completo.

5.4 Easy Life

La empresa Easy LIfe surge a partir de la operadora de telecomunicaciones AWA (Accesos Web Alternativos). Disponen de un equipo multidisciplinar para la investigación y desarrollo de hardware y software dedicado a servicios domóticos y de Hogar Digital e inmótica.

FIGURA 2

Han desarrollado un sistema propio de control por voz para controlar dispositivos con tecnología KNX. Disponen de un interfaz con pantalla táctil empotrable de 4,7'' que realiza la parte de reconocimiento y síntesis de voz además de interfaz con el BUS domótico instalado (puede utilizarse KNX o X10). Esta interfaz es llamada Pasarela Residencial. La pasarela dispone de toda la conectividad necesaria: WIFI, GSM/GPRS, TCP/IP y teléfono. En un principio su esfuerzo está centrado en el desarrollo de aplicaciones de voz para personas discapacitadas, si bien es cierto que el sistema tiene potencial suficiente para comercializarse en un futuro para todo tipo de aplicaciones de Hogar Digital.

El sistema permite transformar las órdenes dadas a su sistema en lenguaje natural en mandatos para poner en marcha diversos dispositivos domóticos: encender la luz, bajar las persianas, marcar el teléfono o efectuar cualquier otra tarea que esté "domotizada". Dispone de voces sintetizadas masculinas y femeninas, cada una responde a su nombre, siendo esta la palabra de atención necesaria para iniciar la comunicación. El sistema puede comandarse por voz también mediante llamadas telefónicas o mediante Internet. Dispone de un módulo GSM que a su vez puede generar mensajes de texto y enviarlos a los móviles deseados. Una aplicación novedosa e interesante es que el usuario puede dictar al sistema una lista de la compra o simples recordatorios mediante la voz. El sistema registra esa lista o recordatorios. Si el usuario desde fuera de casa requiere (por llamada telefónica de voz) la lista o las notas tomadas previamente, el sistema envía al móvil un mensaje de texto con la lista de la compra.

La captación se realiza mediante micrófonos instalados en las paredes de la vivienda de forma fija. La transmisión de las señales la realizan de forma analógica. El sistema de control cuenta con una tarjeta de sonido multicanal que permite la gestión de varias entradas y salidas independientes. El sistema localiza al usuario al ser capaz de determinar qué micrófono que está captando la voz.

Disponen también de unos dispositivos de tipo colgante o pulsera, indicados para personas con problemas de movilidad o incluso para el control de errantes, que detectan en qué habitación se encuentra mediante el uso de la tecnología inalámbrica Zigbee. Detectan mediante el uso de acelerómetros integrados en el dispositivo, si el usuario ha sufrido una caída o desvanecimiento. Están desarrollando sistemas mixtos de captación de voz y detección de movimiento/verticalidad que se puedan instalar en sillas de ruedas si el usuario va a ser una persona minusválida. De esta forma se puede aumentar la autonomía de funcionamiento de los sistemas inalámbricos.

Adicionalmente disponen de un Media Center con todas las funcionalidades típicas de estos sistemas que permite su integración con el sistema de control por voz.

Ventajas:

- Integración con distintos sistemas domóticos (KNX, X10)
- Reconocimiento de voz en lenguaje natural.
- Seguimiento del usuario por la vivienda.
- Humanización del sistema con palabra de atención y diferentes voces de respuesta.
- Interfaz multimodal y con alta conectividad.

Inconvenientes:

- Instalación compleja. Se debe instalar por un lado el bus del sistema KNX o X10, y por otro lado el cableado de audio para micrófonos y altavoces.
- Posibles problemas de interferencias en el audio al transportarse las señales en formato analógico.

5.5 Indisys

Intelligent Dialogue Systems S.L., Indisys, surge tras más de 12 años de investigación en el campo del Procesamiento del Lenguaje Natural y la participación de sus fundadores en proyectos europeos, nacionales y autonómicos, desarrollado una tecnología punta en Inteligencia Artificial, Procesamiento del Lenguaje Natural y Ciencia Cognitiva, con el Diseño Centrado en el Usuario (UCD) y una gran experiencia en las necesidades de clientes y usuarios finales.

Todos los productos y servicios están orientados en el diálogo inteligente y el lenguaje natural. De esta forma el sistema es capaz de interpretar el significado de las frases y ser capaz de anticipar o realizar preguntas para recavar mayor información sobre el asunto que pretende conocer. Para el control de sistemas domóticos han desarrollado una aplicación denominada Mayordomos Virtuales., que permite al usuario controlar los electrodomésticos y otros dispositivos domóticos de su casa u oficina usando como interfaz la voz. Disponen de interfaces para control de sistemas KNX y X10. La aplicación corre sobre un sistema central. Atiende a palabra de control según el apelativo o nombre que se elija. Incorpora un interfaz 3D dinámico, generado a partir de una ontología (esquema conceptual del lenguaje y sus reglas), que sin entrenamiento previo, responde a las peticiones del usuario. Cuando el sistema está inactivo el avatar permanece dormido.

Como ejemplo se puede destacar el siguiente:

"*Ambrosio, ¿Cuántas luces hay encendidas?".*

El sistema reconoce las palabras e interpreta su significado. Al estar conectado con el bus domótico, tiene información sobre los dispositivos conectados y es capaz de saber el número de luces encendidas.

"Hay dos luces encendidas".

El potencial de los sistemas desarrollados por Indisys se basa en el uso de un cerebro artificial denominado Loquaz que emula la inteligencia humana siendo capaz de llevar a cabo un diálogo inteligente y procesar la información en relación al contexto, el histórico de la conversación y realizar un análisis sintáctico y semántico profundo. De esta forma puede

elaborar respuestas naturales o solicitar más información al usuario o buscar información en fuentes externas. Se incluye una captura de pantalla del sistema de mayordomo virtual en la que puede apreciarse el avatar situado en la esquina superior izquierda. Cuando el sistema contesta la sensación es de estar hablando con el avatar.

Ventajas:

- Integración con distintos sistemas domóticos (KNX, X10)
- Reconocimiento de voz en lenguaje natural y emulación de inteligencia humana.
- Seguimiento del usuario por la vivienda.
- Interfaz multimodal.
- Humanización del sistema con palabra de atención, avatar y voces de respuesta personalizables.

Inconvenientes:

- Todavía en fase de desarrollo sin comercialización
- Posibles problemas de interferencias en el audio al transportarse las señales en formato analógico.

5.6 Comparativa de los sistemas comerciales analizados

De los sistemas comerciales descritos en los puntos anteriores, dos de ellos, Fagor y Proinssa son los que se comercializan a fecha de redacción de este proyecto y están funcionando en entornos reales. Destacar que son los sistemas más simples ya que basan su funcionamiento en el reconocimiento de comandos mediante menús. En ambos casos la desventaja es que el reconocedor es un sistema portátil (pulsera, colgante o mando) que requiere la recarga de batería periódicamente. Como ventaja señalar la robustez frente al ruido debido a la portabilidad del dispositivo, que estará siempre cerca del locutor.

Los otros 3 sistemas analizados (de Personica, de Easy Life y de Indisys) están basados en el reconocimiento de voz en lenguaje natural, siendo unos sistemas mucho más complejos. A día de hoy no hay ninguno de ellos instalado y funcionando en una vivienda real. Es cierto que pueden probarse en sus versiones de demo pero aún no están disponibles como producto definitivo para su venta. La ventaja común de todos ellos es su desarrollo tecnológico avanzado, muy orientado hacia la máximas prestaciones en el reconocimiento y síntesis de voz además de la compatibilización con otros protocolos domóticos (KNX y X10 en Indisys y Easy Life y Upnp en el caso de Personica).

En el caso de Easy Life e Indisys el sistema central realiza las tareas de reconocimiento y síntesis mientras que en el caso de Personica son los elementos captadores y emisores los que concentran de forma independiente la inteligencia. En el caso de Personica el BUS utilizado (red LAN o WAN) no transporta las señales de audio, solo lleva señales de control hacia y desde los elementos domóticos. En los otros dos casos, como se indicaba en el apartado de topologías, el transporte de audio se realiza en formato analógico, desde la unidad central hasta los captadores y emisores con el posible problema de interferencias en las líneas de audio. Los tres sistemas incorporan el concepto de interfaz multimodal permitiendo al usuario realizar de formas diferentes la misma acción. También todos ellos hacen uso de una palabra de atención para llamar al sistema y humanizarlo. En el caso de Indisys van un paso más allá, y además de disponer de una voz que imprime emociones según la conversación, han querido ponerle cara mediante el uso de un avatar para dar una sensación mayor de personificación del sistema.

Una de las carencias encontradas en estos sistemas más avanzados, y por tanto más complejos, es que no disponen de indicaciones y consideraciones técnicas para la instalación e integración con los sistemas de control domótico, precisamente el punto hacia el que se ha desarrollado este trabajo de fin de máster.

6 ESTIMACIÓN CUALITATIVA

Desde un punto de vista económico, es claro que a mayor tecnología y desarrollo, mayor coste del producto final. De los sistemas analizados, siempre teniendo en cuenta el sistema de control por voz y no el resto de la instalación domótica asociada, el menor precio se corresponde con los sistemas de control por comandos de Fagor y Proinssa que va desde los 800 € en el primer caso a los 4000 € en el mando SICARE BASIC. Hay que comentar que la solución de Proinssa está orientada a entornos hospitalarios y a personas con discapacidad. Sus tarifas están en la línea de dicho mercado, en el que los precios suelen ser altos y no tienen que ver con los de la electrónica de consumo. En ambos casos la aplicación no realiza síntesis de voz.

Para los sistemas que disponen de una inteligencia mayor y realizan tanto reconocimiento como síntesis de voz permitiendo que el usuario hable en lenguaje natural se puede hablar de precios para el sistema completo de voz (hardware + software) que varían entre 8.000 y 15.000 €.

Es de esperar, que como ocurre con la electrónica de consumo, los precios se vean abaratados en el momento que múltiples fabricantes entren en el mercado del control por voz y el Hogar Digital, a la vez que los usuarios demanden dichas aplicaciones. A día de hoy puede considerarse una tecnología cara si se tiene en cuenta que a esos precios hay que sumar el resto de la instalación de Hogar Digital.

7 PROPUESTAS, CONCLUSIONES Y POSIBLES AREAS DE TRABAJO FUTURAS

Teniendo en cuenta todo lo anteriormente estudiado, una propuesta de sistema “ideal” para el control por voz del Hogar Digital, orientado a todo tipo de vivienda y usuario sería:

- Sistema centralizado, con una unidad de procesado que realice la síntesis de voz y haga de interfaz con el protocolo domótico usado. Este sistema puede ser un ordenador industrial con un microprocesador de doble núcleo y al menos 3 GB de memoria RAM y 40 GB de disco duro. Debe contar con una conexión para red LAN.

- Transmisión de datos de audio mediante red LAN, WAN o PLC, es decir compatibilidad TCP/IP.

- Sistema captador formado por un hardware que realiza parte del preprocesado de audio: filtrado paso banda, ajuste de ganancia, reducción de ruido y cancelación parcial de ecos. Debe contar con al menos dos entradas para dos micrófonos

omnidireccionales instalados en cada estancia. El audio es empaquetado siguiendo las directrices que establece el protocolo UPnP.

- Sistema emisor formado por un hardware que amplifica y reproduce a través de un altavoz el audio que envía la unidad central en forma de streaming siguiendo igualmente las directrices del protocolo UPnP.

El sistema debe ser capaz de reconocer al hablante y en función de sus preferencias y permisos actuar de forma diferente. Deberá permitir las comunicaciones simultáneas desde puntos diferentes de la vivienda o incluso desde fuera de ella. Estará interconectado al sistema de control domótico mediante un interfaz y permitirá el uso multimodal. Para ello el usuario dispondrá de varios interfaces: pantallas táctiles, ordenadores o PDAs con los que podrá realizar las acciones de control y gestión al igual que lo realiza con la voz.

Después de realizar un estudio del estado del arte, de las aplicaciones comerciales y de futuros desarrollos centrados en el control por voz, la principal conclusión es que, aún habiendo tecnología e implementaciones diferentes, no hay una aplicación definitiva que aproveche las ventajas de todas ellas y elimine los puntos débiles de cada una. Tampoco se ha encontrado ningún estudio que trate sobre la integración con las tecnologías de Hogar Digital existentes, como este trabajo ha realizado. Es cierto que estamos en estos momentos, en el inicio de la comercialización de este tipo de tecnología orientada al sector residencial, y que al ir de la mano del Hogar Digital, se esta ralentizando por la dificultad de encontrar un punto en común entre fabricantes y empresas de los sectores eléctrico y de telecomunicaciones.

No hay discusión posible sobre las ventajas de los sistemas de control por voz. Suponen una gran mejora en la calidad de vida de personas con minusvalías físicas o de personas con problemas de movilidad. En este sentido ya hay nichos de mercado que permiten la comercialización de aplicaciones de control de voz mediante comandos que funcionan correctamente para el control de automatismos en viviendas y hospitales. En cualquier caso no debe pensarse el control por voz en el hogar como una aplicación de nicho. La realidad es que introduce una mejora sustancial en la calidad de vida, en la seguridad y en el confort del hogar para todas las personas que lo habitan.

Un detalle técnico comprobado es la imposibilidad de utilización, por falta de ancho de banda, de los buses de los sistemas de control domótico comerciales (exceptuando las redes LAN, WAN y PLC) del bus de transmisión de señales del control por voz.

Como detalle de diseño previo a la instalación, señalar el estudio de la ubicación de los captadores de voz en función de las fuentes de ruido existentes en una vivienda. Algunas de ellas son conocidas y pueden preverse: ruido de electrodomésticos y dispositivos de audio y video, y otros pueden ser más difíciles de anticipar: ruidos externos a la vivienda, conversaciones no deseadas dentro de la vivienda, mascotas, etc. Una adecuada elección de la ubicación de los elementos captadores aumenta considerablemente la eficiencia en el reconocimiento.

Centrando la atención en los sistemas de Hogar Digital analizados y concretamente en el control por voz en lenguaje natural, hay que destacar el trabajo que está realizando Indisys, sin duda una empresa pionera en España, que dispone de un sistema de diálogo inteligente de muy alto nivel. También mencionar el esfuerzo realizado en la integración y comercialización de un producto sólido de Hogar Digital, igualmente con control por voz en lenguaje natural por parte de Easy Life que seguro que acabará siendo un referente para muchas empresas de nueva formación.

En los estudios de mercado de productos tecnológicos es muy importante valorar el rechazo inicial que puede existir por parte del usuario al introducir una nueva tecnología. La clave para evitarlo está en la naturalidad del interfaz. Cuando más se acerque la máquina al hombre y no al contrario, mas probabilidad de éxito hay. En ese sentido, empresas como Loquendo están desarrollando interfaces de síntesis de voz cada vez más reales, recreando emociones, acentos, permitiendo expresiones y frases hechas y en definitiva, todo lo que caracteriza al lenguaje natural. En este punto se ha comprobado que la personalización del sistema, identificándolo con un nombre propio, e incluso asociándole una imagen del tipo avatar, acerca la interacción hombre-máquina y la hace más natural para el usuario. El objetivo final perseguido es que usuario no perciba conscientemente la presencia del interfaz y realice una interacción en forma de conversación natural como si el sistema fuera otra persona. Otra gran ventaja, debida fundamentalmente a la mejora de las técnicas de procesado de voz, es el reconocimiento biométrico del hablante, que da un grado adicional de humanidad al sistema haciendo que reconozca a diferentes usuarios en la vivienda de igual forma que lo hacemos nosotros normalmente según el timbre de cada persona.

Por último y por realizarse este estudio dentro del master en Hogar Digital, es necesario reconocer la necesidad de una campaña informativa, realizada por los implicados en el sector, para poner en conocimiento del usuario final las ventajas (ahorro, confort y seguridad) que suponen a largo plazo la inclusión de estos sistemas en el hogar. De esta forma el usuario será capaz de demandar a vendedores y promotores de nueva vivienda estas calidades, que de ninguna forma deben entenderse como opcionales o de lujo sino como necesidades básicas de las viviendas de los años que vienen.

Como posibles áreas de trabajo futuras se indican algunas ya señaladas en apartados anteriores y otras adicionales derivadas de la propuesta anterior:

- *Estudio de la forma óptima de encapsular y compatibilizar el sistema de control por voz con el sistema BUSing funcionando sobre RS485 y sus limitaciones de utilización.*

- *Estudio económico en la forma de transmitir el audio de síntesis de voz desde la central hasta los elementos emisores: la primera con amplificadores distribuidos, uno instalado junto a cada altavoz; y la segunda, con amplificadores incluidos en las salidas del sistema central.*

- *Estudio económico completo: coste del sistema de control por voz, coste del sistema domótico, proyecto de ingeniería domótica y costes de instalación (cableado y*

mano de obra) comparando un sistema que dispone de dispositivos que realizan de forma autónoma el reconocimiento y la síntesis (caso del sistema de Personica) , y otro sistema del tipo centralizado, con dispositivos con menor carga computacional(caso de Easy Life o Indisys).

- *Estudio y análisis de la instalación de los elementos captadores en diferentes estancias, valorando el impacto visual y las pérdidas en eficiencia del reconocedor en función de la ubicación y de las características de la estancia.*

- *Valoración del coste de fabricación de elementos separados para la captación y la emisión del audio o la unión de ambos en un solo dispositivo. Estudio del tamaño y la forma óptima para un menor rechazo de los usuarios.*

- *Línea de investigación del estudio de interfaces de voz diseñados para el uso de personas con discapacidades físicas (problemas en el habla o en la audición) o discapacidades intelectuales.*

- *Estudio psicológico de las ventajas e inconvenientes de disponer de un sistema "tutor virtual" que pueda controlar y estimular el estudio y el ocio de los niños dentro de la vivienda basado en un sistema de voz con diálogo inteligente.*

8 REFERENCIAS Y BIBLIOGRAFÍA

Muchas referencias utilizadas se encuentran ya incluidas en el documento en forma de hipervínculos. No obstante se incluyen en esta sección algunas referencias más técnicas para que el lector pueda consultar las fuentes de estudio teóricas.

- WOZ experiments in Multimodal Dialogue Systems
Pilar Manchón, Guillermo Pérez & Gabriel Amores (2005) Proceedings of the ninth workshop on the semantics and pragmatics of dialogue, 131-135. Nancy, France. June, 2005.

- Knowledge-based Reference Resolution for Dialogue Management in a Home Domain Environment
J. F. Quesada & J. G. Amores (2002) Johan Bos, Mary Ellen and Colin Matheson, eds. Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (Edilog). 4-6 september 2002. pp 149-154

- A Dynamic Approach for the Specification and Reasoning of Discourse Knowledge in Man-Machine Dialogue Systems
G. Fernández, G. Amores & J. F. Quesada (2000) A. Nepomuceno, J.F. Quesada, F.J. Salguero (eds.) Lógica, Lenguaje e Información. Actas de las Primeras Jornadas sobre Lógica y Lenguaje, pp. 87-96.

- MIMUS: A Multimodal and Multilingual Dialogue System for the Home Domain. J. Gabriel Amores, Guillermo Pérez & Pilar Manchón. (2007) Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, pages. 1-4. ISBN: 978-1-932432-87-9. 23-30 June 2007.

- Semi-Automated Testing of Real World Applications in Non-Menu-Based Dialogue Systems.
Pilar Manchón, Guillermo Pérez, Gabriel Amores, Jesús González. (2007) Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pages 181-182. Trento, Italy, 30 May - 1 June 2007.

- A Multimodal Architecture for Home Control by Disabled Users
Guillermo Pérez, Gabriel Amores & Pilar Manchón. (2006) Proceedings of IEEE/ACL Workshop on Spoken Language Technology (SLT), Aruba. December 2006.

- Interfaz Robusta para el Reconocimiento de Comandos Hablados y remodelado de un Procesador Soporte de la misma utilizando Metodologías de Codiseño",
- Proyecto de Investigación financiado por el Programa Nacional para las Tecnologías de la Información y las Comunicaciones. Referencia: TIC97-1011. Director: Victoria Rodellar Biarge.

- Diseño de un corpus par una base de síntesis de voz.
Ignacio Hernández, Asunción Moreno. Universidad Politécnica de Cataluña

- Verificación de hablante basado en Dynamic Time Warp.
Lácides Antonio Ripoll Solano.

- Desarrollo de un Segmentador Automátivo de voz mediante Modelos Ocultos de Markov.
Luis Fernando D'Haro. Docente Universidad Autónoma de Occidente.

- Modelos ocultos de Markov para el reconocimiento del habla.
d.milone@ieee.org

- Introducción a los modelos ocultos de Markov
Luis Miguel Bergasa. Departamento de Electrónica. Universidad de Alcalá.

- El Filtro de Wiener.
Miguel Angel Lagunas.2.003

- Interfaces Multimodales.
Publicaciones Telefónica.