EDECÁN: sistEma de Diálogo multidominio con adaptación al contExto aCústico y de AplicacióN

Eduardo Lleida[1], Encarna Segarra[2], María Inés Torres[3], Javier Macias[4]

[1] Universidad de Zaragoza (uz@edecan.es)
[2]Universitad Politécnica de Valencia(upv@edecan.es)
[3]Universidad del País Vasco (ehu@edecan.es)
[4]Universidad Politécnica de Madrid (upm@edecan.es)

RESUMEN

El proyecto EDECÁN[1] tiene como objetivo aumentar la robustez de un sistema de diálogo de habla espontánea a través del desarrollo de tecnologías para la adaptación y personalización del mismo a los distintos contextos acústicos y de aplicación en los que pueda encontrarse. El concepto de contexto acústico engloba todos los elementos que influyen, en mayor o menor medida, en la señal que capta/n el/los micrófono/s que conforma/n la entrada del sistema de diálogo. Estos elementos dependen tanto del usuario como del entorno físico que lo rodea. Por otro lado, el contexto de aplicación hace referencia a la estructura semántica de los dominios en los que se desarrolla el diálogo.

Los objetivos de EDECÁN, implican la necesidad de desarrollar estrategias que permitan caracterizar las condiciones de funcionamiento del sistema de diálogo (condiciones acústicas, tipo de habla utilizada, contexto semántico, tipo de usuario ...) y definir e implementar técnicas de adaptación a tales condiciones. La incorporación de técnicas de adaptación y personalización dinamizan el funcionamiento de un sistema de diálogo, lo que obliga a adaptar a esta nueva situación las estrategias de evaluación y medida de usabilidad conocidas. Por otro lado, este proyecto aborda la extensión del sistema de diálogo a nuevos contextos de aplicación, posibilitando al usuario la consecución de múltiples objetivos en el transcurso del diálogo.

EDECÁN es un proyecto coordinado en el que participan investigadores de la Universidad Politécnica de Valencia, Universidad del País Vasco, Universidad Politécnica de Madrid y Universidad de Zaragoza.

1. INTRODUCCIÓN

Los avances recientes en las Tecnologías de la Información y las Comunicaciones, han tenido un gran impacto en el modo en que vivimos, trabajamos e interactuamos con nuestro entorno personal y profesional. Estas tecnologías están permitiendo desarrollar redes distribuidas de sistemas que proporcionan información, comunicación y entretenimiento. En este contexto, una visión futurista de la Sociedad de la Información enfatiza el desarrollo de entornos en los que las personas interactúan de forma transparente con multitud de dispositivos interconectados para desarrollar actividades de la vida diaria. Este planteamiento se materializa, a grandes rasgos, en un individuo rodeado de interfaces inteligentes e intuitivas que se encuentran integradas en elementos de la vida cotidiana, todo esto en un entorno que sea capaz de reconocer y responder a la presencia y necesidades de diferentes individuos, de una forma personalizada, discreta e imperceptible salvo a través de los resultados. El entorno mencionado, bautizado como "ambiente inteligente", está donde el individuo se encuentre y responde a sus necesidades de una forma natural, no limitándose a un único lugar físico sino que comprende a todos ellos, la casa, el coche, el lugar de trabajo, etc. Uno de los retos más apasionantes que pone sobre la mesa esta visión es el desarrollo y la investigación en sistemas de diálogo robustos como interfaz entre las personas y el "ambiente". Un valor añadido de esta investigación es su contribución a la incorporación a la Sociedad de la Información, de la población no habituada al manejo de las nuevas tecnologías al desarrollar una interfaz "intuitiva" entre las personas y los recursos que la Sociedad de la Información pone a su disposición.

Sin embargo, hoy en día, los sistemas de diálogo son más o menos aceptados y útiles cuando trabajan en entornos semánticos muy restringidos, con condiciones acústicas favorables y con usuarios colaboradores. Los portales de voz son un ejemplo ya que éstos están preparados para unas condiciones acústicas particulares (canal telefónico fijo y/o móvil) y aplicaciones muy concretas. Por todo ello, es necesario todavía dedicar un gran esfuerzo de investigación para dotar, a todos los componentes de un sistema de diálogo, de robustez y capacidad de adaptación frente a los diversos contextos y necesidades propias de cada usuario en los que puede encontrarse el sistema.

En un sistema de diálogo se encuentran sistemas de distinta complejidad que deben reconocer las palabras pronunciadas, comprender su significado, gestionar el diálogo (incluyendo la información contextual, manejo de errores y acceso a la aplicación final) y generar finalmente la respuesta oral. Además de estas funciones básicas, el sistema puede incorporar nuevas funcionalidades como el modelado del usuario, la adaptación al entorno, o la inclusión de otras formas de entrada/salida como texto, lectura de labios, gráficos o sistemas táctiles, conformando así un sistema multimodal que robustece, facilita y mejora la interacción de los usuarios, incluso con algún tipo de minusvalía.

[1] Este proyecto ha sido subvencionado por el Gobierno Español mediante el proyecto coordinado TIN2005-08660-C04.
EDECÁN (Del fr. aide de camp). Ayudante de campo. Auxiliar, acompañante.(Diccionario 2001 Real Academia Española)

Desde el punto de vista de la complejidad del diálogo, Allen et al. [1] definen 5 niveles de complejidad según la técnica utilizada para representar los actos de diálogo, tal y como se muestra en la figura 1.

Técnica Empleada	Ejemplo de Tarea	Complejidad de la Tarea	Fenómenos del Diálogo que Gestiona
Diagramas de Estados Finitos	Marcación telefónica	Menos Compleja	Respuesta a preguntas del usuario
Basado en Frames	Información de llegadas y Salidas de Trenes		Respuesta a preguntas del usuario, clarificaciones simples del sistema
Conjunto de Contexto	Agente de Viajes	Más Compleja	Conmutación entre dominios semánticos predeterminados
Modelos basados en planes	Consultor de diseño de cocinas		Estructuras de dominios semánticos generadas dinámicamente, subdiálogos de negociación colaborativa
Modelos basados en agentes	Gestión de apoyo a catástrofes.		Diferentes modalidades.

Figura 1.Niveles de complejidad en un sistema de diálogo [1]

Bajo este contexto, el proyecto EDECÁN propone como objetivo aumentar la robustez del sistema de diálogo a través del desarrollo de tecnologías para la adaptación y personalización del mismo a los distintos contextos acústicos y de aplicación en los que pueda encontrarse. Este objetivo implica la necesidad de desarrollar estrategias que permitan caracterizar las condiciones de funcionamiento del sistema de diálogo (condiciones acústicas, tipo de habla, dominio de aplicación, tipo de usuario, ...), definir e implementar técnicas de adaptación a tales condiciones, extender el sistema de diálogo a nuevos dominios de aplicación, y estudiar la evaluación y usabilidad de los sistemas de diálogo.

Desde el punto de vista del diálogo, EDECÁN es una evolución de dos proyectos anteriores del equipo investigador: DIHANA (TIC2002-04103-C03) y SAITE (FEDER 2FD1997-1062-C02-01). El proyecto DIHANA tuvo como objetivo la profundización en aspectos metodológicos fundamentales sobre la interfaz de audio y modelado acústico, tratamiento del habla espontánea, modelado del lenguaje, comprensión y diálogo, mientras que SAITE se centró en el estudio de estrategias avanzadas de diálogo multiobjetivo. En ambos casos la aplicación de diálogo estaba restringida a un servicio de información de precios y horarios de tren. Dentro de la taxonomía definida en la figura 1, DIHANA y SAITE, como muchos de los sistemas de diálogo construidos hoy en día [1], se pueden clasificar como sistemas de baja complejidad basados en “frames” con iniciativa mixta.

Sin embargo, EDECÁN supone un paso más en dicha taxonomía al aumentar la complejidad a sistemas basados en múltiples dominios, cada dominio representado por una serie de contextos, cada uno de ellos basado a su vez en “frames”. Este aumento en la complejidad del sistema de diálogo conlleva nuevos retos como son la identificación de los cambios de contexto y dominio que el usuario puede realizar (por ejemplo, cuando el usuario vuelve a un contexto anterior y quiere cambiar algún detalle). Además, EDECÁN propone la personalización, a un usuario particular, del sistema de diálogo tanto a nivel del contexto acústico como de aplicación. La personalización conlleva la identificación de usuario (o tipo de usuario) y del entorno acústico, la definición de perfiles de usuario, etc. Por todo ello, EDECÁN tiene también como objetivo evolucionar el diseño de la arquitectura del sistema de diálogo DIHANA para facilitar la adopción e incorporación de las nuevas tecnologías que se desarrollen.

2. LOS SISTEMAS DE DIÁLOGO HABLADO: NUEVOS RETOS RELACIONADOS CON LOS CONTEXTOS ACÚSTICO Y DE APLICACIÓN

Los objetivos en el desarrollo de sistemas de interacción hombre-máquina han evolucionado convirtiéndose en más realistas (a la vez que más ambiciosos) conforme se han desarrollado las tecnologías necesarias. Sin embargo, aunque bajo ciertas condiciones, el modelado de cada una de las áreas que participan en los procesos de interacción oral pueden proporcionar buenos resultados (modelos acústicos, lingüísticos, o de diálogo), tanto el problema de la implantación de estos sistemas en ambientes reales como la adaptación a nuevas tareas son un reto que va a exigir importantes esfuerzos: reconocimiento del habla espontánea con usuarios no “entrenados” y en ambientes ruidosos, modelado del lenguaje y semántico con una amplia cobertura, o la gestión del diálogo dando máxima libertad al usuario.

En los últimos años ha habido importantes avances en el desarrollo de sistemas de diálogo hablado, tanto en la línea de servicios de información (donde la interacción con la máquina requiere una cierta complejidad para conseguir la información deseada), como en servicios donde la interacción es más sencilla pero las condiciones acústicas dificulten mucho la tarea.

2.1. Sistemas de diálogo y contexto de aplicación

Por contexto de aplicación nos referimos a la estructura semántica de los dominios en los que se desarrolla el diálogo. Esta estructura incluye: los conceptos semánticos y las relaciones entre ellos (ontología), el conjunto de reglas (diseñadas o aprendidas) que relacionan conceptos con expresiones del lenguaje natural, y por último, un modelo para la gestión eficiente de la interacción persona-máquina. Una de las características más importante de un sistema de diálogo es que tenga capacidad de adaptarse a nuevas tareas, a diferentes tipos de usuario o de idioma, o a cambios en las condiciones del contexto acústico. Por ello se han dedicado muchos esfuerzos al diseño de arquitecturas que permitan fácilmente la adaptación a nuevos dominios. Tal es el caso de la arquitectura GALAXY [2] del MIT bajo la cual se han desarrollado los sistemas JUPITER (información meteorológica), VOYAGER, (información urbana), u ORION (asistente personal) y que ha servido de plataforma para el proyecto DARPA COMMUNICATOR, http://communicator.sourceforge.net . A nivel nacional, los proyectos BASURDE (TI98-0423-C06) y DIHANA han definido una arquitectura, similar a GALAXY, que permite la integración de nuevos componentes al sistema de diálogo.

Aparte del gestor de diálogo (módulo fundamental en cualquier sistema de diálogo) el módulo de comprensión de lenguaje natural juega también un papel primordial. Aunque tradicionalmente la construcción de este módulo se ha basado en la definición manual de reglas semánticas para la detección

de palabras clave con las que rellenar un marco (frame), también se han desarrollado otras aproximaciones basadas en el uso de modelos estocásticos: el BNN-HUM [3], el AT&T-CHRONUS [4] y el LIMSI-ARISE [5] son algunos ejemplos del uso de modelos ocultos de Markov y N-gramas para modelar estocásticamente el proceso de comprensión. Hay también otras aproximaciones estadísticas basadas en clasificación, traducción, y técnicas de inferencia gramatical. Estas aproximaciones estocásticas entienden el proceso de comprensión como un problema de traducción de la frase de entrada a una determinada representación semántica del significado de dicha frase. Una de las principales ventajas de la aproximación estocástica es que permite aplicar técnicas de aprendizaje automático a partir de datos reales modelando las posibles fuentes de error.

La existencia de varios contextos de aplicación implica un diseño del sistema mucho más complejo, en el que se debe identificar el contexto apropiado para dar respuesta al usuario en cada momento, dando la posibilidad de cambiar de contexto a medida que se vaya desarrollando el diálogo [1]. Un reto importante es detectar los casos en los que el usuario quiere volver a un contexto anterior y modificar algún detalle.

La variedad de dominios exige también el diseño de varios modelos de lenguaje con diferente complejidad así como de potentes mecanismos de integración y coordinación. Por otra parte, las disfluencias acústicas debidas al habla espontánea que aparecen al cambiar de dominio, también pueden ser objeto de tratamiento a nivel del modelo lingüístico. Se puede modelar el lenguaje incluyendo específicamente estos fenómenos como eventos del léxico o bien considerarlos un ruido del canal y tratar de filtrarlos. El tratamiento de las disfluencias sintácticas (reformulaciones, falsos comienzos, etc.) mediante un modelo de lenguaje específico suministra información útil al módulo de comprensión sobre posibles borrados, inserciones o sustituciones que pueden alterar la interpretación de las intervenciones del usuario.

2.2. Sistemas de diálogo y contexto acústico

El contexto acústico lo podemos definir como el conjunto de elementos que influyen, en mayor o menor medida, en la señal captada por el/los micrófono/s que conforma/n la entrada del sistema de diálogo. Estos elementos se dividen en dos grandes grupos: aquellos que dependen del usuario y los que dependen del entorno físico que le rodea.

2.2.1. Adaptación a los condicionantes debidos al usuario

En este caso, los principales condicionantes son la edad, el sexo, lugar de origen, velocidad de locución, estado emocional, espontaneidad, grado de cooperación, etc. Aunque se ha comprobado que un sistema independiente del locutor puede obtener altas tasas de reconocimiento, éstas sólo son alcanzables en condiciones de habla controlada donde la locución del usuario se adapte al tipo de frase predominante en el corpus utilizado para el aprendizaje. Sin embargo, existe un gran grupo de usuarios para los que, por las características particulares de su voz, estos sistemas presentan unas prestaciones inaceptables. Así, por ejemplo, para usuarios con cierto tipo de minusvalías que lleven asociadas un desorden fisiológico del sistema fonador (disartria), para los cuales la utilización de un sistema de diálogo sería de gran ayuda, los sistemas independientes del locutor son totalmente inservibles. Por otra parte, velocidad de locución, estado emocional y espontaneidad, condicionantes muy ligados entre sí, dotan de gran variabilidad a la señal de voz, influyendo negativamente en las prestaciones de los sistemas de reconocimiento del habla.

Las tecnologías de adaptación y personalización al usuario son las respuestas que actualmente se están dando a los problemas derivados de los condicionantes del usuario. La adaptación al usuario se puede realizar mediante la transformación de los vectores de características extraídos de la señal de voz, o con la transformación de los modelos acústicos partiendo del menor número posible de muestras de voz, o mediante la identificación de tipos de hablantes con características similares para los que se entrenan modelos, acústicos y de lenguaje, específicos. El reconocedor realiza una adaptación rápida tras seleccionar el conjunto de modelos a aplicar en cada caso.

La adaptación al estado emocional del hablante es muy importante en entornos más personales del usuario (hogar, automóvil, trabajo, etc) y con habla espontánea (en el vocabulario, el lenguaje o la semántica). Diversas estrategias para mejorar las tasas de reconocimiento son: clasificación y modelado, nuevos parámetros de análisis o técnicas de restauración de la señal de voz. En los sistemas de diálogo actuales se reducen las prestaciones del reconocedor al tratar con habla espontánea. Este hecho se debe al elevado número de “disfluencias” que aparecen, incluso en diálogos usuario-máquina. El tratamiento acústico específico de las disfluencias debidas al habla espontánea es necesario para frenar esta reducción. Por otra parte, en comunidades bi o multilingües es posible adaptar los modelos acústicos y/o lingüísticos a las preferencias del usuario.

Los cambios en la velocidad de locución conlleva la modificación de la duración de los sonidos, la alteración espectral de los mismos, y cambios profundos en los efectos de coarticulación. La modificación de la duración se ha atacado proponiendo diversos modelos de duración, mientras que la alteración espectral y los efectos de coarticulación se puede englobar dentro de las técnicas genéricas de adaptación al locutor y modelado acústico.

El trabajo para adaptar los sistemas de reconocimiento automático del habla a las alteraciones fisiológicas del sistema de generación de la voz es muy complejo. Aún así, se han realizado trabajos que optan por un replanteamiento del modelado acústico, la transformación de la voz alterada a una patrón o la adaptación al locutor. Los resultados obtenidos demuestran que es necesario adaptar y personalizar el sistema a cada hablante específico dada la gran variabilidad en la voz. Especialmente interesante para estos casos es el uso complementario de información multimodal. El trabajo en la adaptación acústica al usuario es un auténtico reto por la escasez de corpus especializados para el estudio y modelado de la variabilidad de voz debida a los condicionantes dinámicos (emoción, desordenes fisiológicos, …), y por la necesidad de desarrollar nuevas tecnologías de adaptación “no estáticas” a este tipo de condicionantes.

La adaptación al usuario involucra también a otros módulos relevantes de un sistema de diálogo como por ejemplo el módulo de comprensión, que utiliza un conjunto de reglas para relacionar expresiones del lenguaje natural con los conceptos semánticos. Cuando se desarrolla un sistema, se plantea un gran abanico de reglas que cubran una gran variedad de expresiones. La personalización de este módulo se centra en seleccionar o particularizar las reglas del sistema a las expresiones que un usuario concreto utiliza con una mayor frecuencia. Esta selección de reglas permite reducir la perplejidad del sistema, mejorando sus prestaciones.

En relación con el gestor de diálogo, que se encarga de dirigir la interacción persona-sistema, la personalización o adaptación al usuario se puede concretar en dos aspectos. La definición de un perfil de usuario con preferencias sobre el uso del sistema de diálogo, puede ser de gran utilidad para reducir el número de turnos de diálogo en interacciones posteriores. Parte de la información necesaria puede obtenerse del perfil de usuario y proponerla sin necesidad de preguntarla (ej: proponer la hora a la que el usuario ha realizado viajes similares con la misma ciudad origen y la misma ciudad destino). El segundo aspecto está relacionado con la variación del ritmo del diálogo según la destreza y/o experiencia del usuario. Usuarios más inexpertos requieren de un ritmo de diálogo más pausado, mientras que un usuario con mayor destreza puede agilizar sensiblemente el proceso.

En los módulos de generación de respuesta y síntesis de voz, la personalización consiste principalmente en adaptar el mensaje al usuario para que asimile la mayor cantidad de información posible. Esta adaptación requiere de una selección inteligente de la expresión a utilizar (más o menos cantidad de información por interacción, diferente tipo de expresión a utilizar: coloquial con tratamiento “de tú” o formal con tratamiento “de usted”), y de la generación acústica posterior. Esta generación acústica puede tener características preferidas por el usuario: voz femenina en lugar de masculina, generación de emociones ante determinadas situaciones, etc. Las voces sintéticas generalmente desarrolladas en la bibliografía están pensadas para servicios por lo general no presenciales (telefónicos) apreciándose más las voces femeninas que las masculinas. La falta de variedad en estas voces sintéticas hace que sean calificadas como monótonas y aburridas en algunas aplicaciones. En un entorno doméstico y personalizado estas voces no resultan adecuadas. A fin de incrementar su amigabilidad, es necesario incorporar variedad emocional y diversidad de locutores, que haga que las voces resulten más agradables.

2.2.2. Adaptación a los condicionantes debidos al entorno físico

Dentro de estos condicionantes podemos encontrar la presencia de ruido aditivo y ruido convolucional sobre la señal de voz. Las prestaciones de un sistema de reconocimiento automático del habla se ven altamente afectadas por el entorno acústico en el cual se encuentra/n el/los micrófono/s que captan la señal de voz. También, un entorno altamente ruidoso provoca la distorsión de la señal de voz por el conocido “efecto Lombard”. En principio, los problemas asociados a la presencia de ambos tipos de ruido pueden paliarse en gran medida con el uso de micrófonos de cercanía unidireccionales (close-talk), situados muy próximos a la boca del locutor. Con ellos, se garantiza una alta relación señal a ruido y una muy baja proporción de ruido convolucional. Sin embargo, la utilización de este tipo de sensores no es posible en la mayor parte de las aplicaciones. En esos casos, el reconocimiento debe realizarse con señales de voz captadas por micrófono/s situado/s a cierta distancia del usuario, donde la relación señal a ruido se degrada considerablemente y empieza a notarse la influencia de la reverberación sobre la señal de voz captada (ruido convolucional). En este caso, distinguimos dos situaciones: una en la que la voz se sitúa en campo directo y el ruido está en campo difuso, y otra, que en los últimos años ha merecido un especial interés, en la que tanto la voz como el ruido se sitúan en campo difuso.

Las técnicas a utilizar para paliar los efectos nocivos de la captura de señales en campo lejano son, entre otras: el empleo de múltiples sensores, procesado basado en la percepción humana biaural, técnicas de procesado de señal para realzar la señal de voz (mejorando la relación señal a ruido de la misma), compensación de los parámetros acústicos de entrada (aproximándolos a los parámetros que se obtendrían con un micrófono de proximidad) o la compensación de los modelos acústicos (obtenidos mediante señal limpia de ruido de fondo y captada en un entorno anecoico asemejándose a los que se obtendrían en el nuevo entorno acústico).

En entornos acústicos especialmente adversos, se está trabajando en la combinación de información procedente de varias fuentes ortogonales para mejorar los sistemas de reconocimiento automático del habla. La información visual es una de esas fuentes no afectadas por el entorno acústico. Se ha demostrado que la incorporación de información visual a los sistemas de reconocimiento automático del habla puede elevar sus prestaciones. Sin embargo, para llevar a cabo una comparativa justa, es preciso contar con datos suficientes tomados en situaciones reales en los que la señal de vídeo también pueda estar afectada por “ruido” visual. Existen tres razones fundamentales por las cuales, la información visual ayuda al reconocimiento del habla: ayuda a localizar al locutor, contiene información segmental del habla que complementa al audio, y puede ayudar a resolver ambigüedades con información adicional sobre el lugar de articulación.

La adaptación al entorno físico también requiere cierto esfuerzo en el módulo de comprensión, el cual está orientado a paliar la pérdida de calidad que sufre el reconocedor de voz cuando se enfrenta a un entorno físico hostil. El principal efecto que se produce en este caso es que ciertas palabras pronunciadas se pierden o se reconocen con error. Para aumentar la robustez de las reglas que relacionan dichas expresiones con los conceptos semánticos hay que dotarlas de mayor flexibilidad para la extracción de los conceptos a pesar de la existencia de errores en la frase reconocida.

En relación con el gestor de diálogo la adaptación al entorno físico se puede concretar en dos puntos: al igual que ocurría con la destreza del usuario, el entorno físico puede hacer que el ritmo de diálogo cambie. El ritmo se ralentiza en condiciones acústicas hostiles para evitar que el usuario se pierda. Por otro lado, el módulo de caracterización acústica puede informar sobre la existencia de determinado tipo de ruido (TV encendida) o situación (micrófono mal colocado) obligando al gestor de diálogo a suspender la interacción e instar al usuario a que resuelva dicha situación (bajar el volumen de la TV o colocarse bien el micrófono) antes de continuar con la interacción.

En relación con el módulo de generación de respuesta, la adaptación se centra en la redacción de las expresiones a pronunciar incluyendo una mayor o menor redundancia de información, según la hostilidad del entorno acústico. En cuanto a la síntesis de voz, la graduación del nivel de volumen y la incorporación de emociones son de gran utilidad para mejorar la percepción.

2.3. Sistemas de diálogo: Evaluación y usabilidad

En USA y en Europa ha habido varias iniciativas para llevar a cabo la evaluación de los sistemas de diálogo y el estudio de su usabilidad. En el contexto americano, hay dos proyectos importantes por su contribución a estos dos aspectos: el proyecto ATIS se centró en la evaluación técnica

y el COMMUNICATOR que ha contribuido principalmente al análisis de la usabilidad de los sistemas. En Europa, varios proyectos han trabajado en la definición de los parámetros de la evaluación de los sistemas de diálogo hablado: el proyecto EAGLES [6] y el DISC [7]. Estos proyectos proponen que la evaluación debe consistir en: una evaluación técnica de los sistemas y sus componentes, una evaluación de la usabilidad de los sistemas y una evaluación de los sistemas sus componentes llevada a cabo por el usuario. En la actualidad, el marco de trabajo más extendido para la evaluación y análisis de usabilidad de sistemas de diálogo es PARADISE [8]. La solución propuesta en PARADISE está muy orientada a sistemas unimodales, ampliándose recientemente a sistemas multimodales: PROMISE [9], pero sin estar completamente desarrollado.

En el proyecto EDECÁN se abordan tres situaciones de evaluación nuevas. En primer lugar un sistema de diálogo multidominio necesita analizar nuevos parámetros de medida que midan aspectos relacionados con el cambio y gestión de varios dominios. En segundo lugar la incorporación de técnicas de adaptación y personalización incorporan cierto dinamismo en el funcionamiento que obligan también a adaptar las estrategias de evaluación y medida de usabilidad conocidas [10]. El tercer aspecto es la evaluación de sistemas con información multimodal, en este caso trabajaremos sobre la propuesta de PROMISE.

3. EL CONSORCIO EDECÁN

El consorcio EDECÁN está constituido por 4 grupos de investigación en Tecnologías del Habla:

EDECÁN-UZ: Grupo de Tecnologías de Comunicaciones, Instituto de Investigación en Ingeniería de Aragón, Universidad de Zaragoza.

EDECÁN- UPV: Grupo de Reconocimiento de Formas, Inteligencia Artificial, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

EDECÁN-EHU: Grupo de Reconocimiento de Formas y Tecnologías del Habla. Departamento de Electricidad y Electrónica. Universidad del País Vasco/Euskal Herriko Unibertsitatea .

EDECÁN-UPM: Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica, Universidad Politécnica de Madrid.

Las diferencias en el origen y especialización de cada uno de los grupos hacen especialmente atractivo el trabajo en conjunto. El grupo EDECÁN-UZ proviene del ámbito de teoría de la señal y las comunicaciones, aportando un conocimiento fundamental para el desarrollo de técnicas de robustez de sistemas de reconocimiento, tanto en técnicas de parametrización de la señal acústica como en el modelado acústico. Los grupos EDECÁN-UPV y EDECÁN-EHU provienen del ámbito del reconocimiento de formas y poseen conocimiento esencial sobre el aprendizaje automático de los modelos. Su presencia permite al consorcio profundizar en el desarrollo de metodologías de aprendizaje automático a partir de muestras, tanto para el tratamiento de fenómenos de habla espontánea como para el diseño de sistemas de modelado de lenguaje, comprensión y diálogo. El grupo EDECÁN-UPV también trabaja en el ámbito del tratamiento del lenguaje natural, en particular en etiquetadores morfosintácticos, desambiguación del sentido de las palabras, reconocimiento de entidades con nombre y su aplicación en recuperación y extracción de información y búsqueda de respuestas. El grupo EDECÁN-UPM tiene una extensa experiencia en el diseño y evaluación de sistemas de diálogo persona-máquina basados en tecnología del habla, con potentes sistemas de reconocimiento, comprensión de habla y conversión de texto a voz de alta calidad.

4. OBJETIVOS DE EDECÁN

EDECÁN es un proyecto de investigación horizontal que necesita de la investigación y desarrollo tanto en tecnologías electrónica y de comunicaciones como en tecnologías informáticas. El proyecto aborda temáticas relacionadas con las técnicas avanzadas de procesado digital de señal y de reconocimiento robusto de voz dentro de los objetivos de robustez frente al contexto acústico. Por otro lado, el proyecto, en su conjunto, aborda el desarrollo de un sistema completo de diálogo multidominio adaptable y personalizable.

Objetivos científicos:

- Optimización y mejora de los componentes de los sistemas de diálogo disponibles:
  - Mejora de la robustez y prestaciones de los sistemas de diálogo disponibles en cada grupo.
  - Diseño e implementación de nuevos componentes (generación de habla y síntesis con emociones e integración básica de información multimodal).
- Búsqueda de mecanismos/técnicas eficientes que mantengan las prestaciones de los sistemas frente a cambios en el contexto acústico (en general: físico, del locutor, etc.). **[Cambio de contexto acústico]**
- Búsqueda de mecanismos/técnicas eficientes que minimicen el esfuerzo invertido para generar un nuevo sistema de diálogo. **[Cambio de contexto semántico]**
- Búsqueda de mecanismos eficientes para que los sistemas exploten la información disponible en un perfil de usuario, con el objetivo de adaptar el comportamiento del sistema a las preferencias / capacidades / características de los usuarios. **[Personalización]**
- Estudio de mecanismos contrastados para llevar a cabo la evaluación de las prestaciones y de la usabilidad. **[Evaluación]**

Objetivos tecnológicos:

- Sobre el corpus
  - Adquisición y trascripción de un corpus específico para la investigación en la adaptación y personalización al contexto acústico y de aplicación.
  - Etiquetado del corpus anterior en términos de los elementos condicionantes del contexto acústico y de aplicación.
- Sobre el prototipo
  - Desarrollo de una plataforma para la integración de las diferentes unidades de los sistemas de diálogo.
  - Construcción de un prototipo de sistema de diálogo que integre los resultados científicos y tecnológicos sobre dos contextos de aplicación definidos por un servicio de información del transporte y un servicio de asistente en el hogar y en el coche.

Por otro lado, dentro de estos objetivos generales del proyecto, cada subproyecto tiene unos objetivos particulares:

El **subproyecto EDECÁN-UZ** se centra en la investigación y desarrollo de tecnologías de adaptación al contexto acústico para los sistemas de reconocimiento

automático de habla. En particular, este subproyecto se plantea como objetivos científicos concretos:

- La investigación en representaciones de la señal de voz robustas para los distintos entornos acústicos
- La identificación de entornos acústicos y la identificación de locutores y tipos de locutores para la personalización y definición de perfiles de usuario y de aplicación
- La adaptación de los modelos acústicos al entorno acústico, locutor, estado emocional, velocidad de locución y alteraciones de la voz.
- El reconocimiento con micrófonos lejanos, especial énfasis se dará a la utilización de agrupaciones de micrófonos y reconocimiento biaural. Se profundizará en el estudio de dos situaciones. Una primera en la que la voz se sitúa en campo directo y el ruido está en campo difuso y una segunda situación, que en los últimos años ha merecido un especial interés, en la que tanto la voz como el ruido se sitúan en campo difuso.
- Nuevos algoritmos de decodificación acústico-fonética que posibiliten la adaptación "on-line" tanto a nivel de modelos acústicos como de parámetros acústicos.

El **subproyecto EDECÁN-UPV** se centra en el área de adaptación de los componentes de comprensión y gestión del diálogo a los cambios de contexto, tanto acústico como de dominio semántico. En particular, este subproyecto se plantea como objetivos científicos concretos:

- El desarrollo de técnicas para mejorar la robustez de las fuentes de conocimiento de los sistemas de reconocimiento, en lo referente a los modelos de lenguaje, modelos de comprensión del habla y en la estrategia y modelos para el gestor de diálogo, frente a los cambios de contexto acústico.
- La búsqueda de técnicas que minimicen el esfuerzo invertido para generar un nuevo sistema de diálogo. Para ello se propone el aprendizaje automático (uso de métodos supervisados y no supervisados) de modelos de lenguaje y de modelos de comprensión del habla, el desarrollo de técnicas de adaptación a nuevas tareas.
- El estudio de metodologías de aprendizaje de la estrategia del diálogo para las diferentes tareas.
- Optimización y mejora de los componentes del sistema de diálogo.

El **subproyecto EDECÁN-EHU** se plantea como objetivos científicos:

- Desarrollo de tratamientos específico para el habla espontánea en los modelos acústicos, en la construcción del léxico y en el modelo de lenguaje.
- Estudio de formulaciones unificadas para la integración ponderada de modelos de estados finitos en el reconocedor.
- Desarrollo de técnicas de adaptación rápida de los modelos acústicos a las características de usuarios desconocidos.
- Desarrollo de modelos de lenguaje específicos para cada dominio y de técnicas de coordinación de los mismos.

Por otra parte se propone incorporar los corpus, modelos y experiencia adquirida en el tratamiento específico del Euskera en reconocimiento automático del habla para poder personalizar los prototipos que se deriven a posibles usuarios vascoparlantes.

El **subproyecto EDECÁN-UPM** se centra en el desarrollo de nuevas tecnologías que permitan mejorar el proceso de desarrollo de Sistemas de Diálogo para su funcionamiento en varios dominios. El trabajo en estas nuevas tecnologías se orienta a tres objetivos principales:

- Adaptación al dominio acústico y de aplicación: adaptación de los modelos acústicos, léxicos y lingüísticos del reconocedor de voz, prestando interés especial al reconocimiento en condiciones de micrófono lejano (localización y enfoque).
- Personalización y adaptación al usuario: adaptación acústica, lingüística y semántica al usuario mediante la incorporación de perfiles de usuario con información de adaptación aplicable en cada uno de los módulos. Una tarea relevante en este punto es el trabajo en identificación de locutor para poder seleccionar el perfil adecuado a aplicar.

Mejora de la gestión de diálogo en un contexto multidominio: aplicación de modelos probabilísticos más potentes (Belief Networks), incorporación de nueva información en la gestión del diálogo (medidas de confianza, información multimodal y perfiles de usuario), desarrollo de herramientas de apoyo al diseño, y mejora de la generación de respuesta (generación de estilos de habla y síntesis con emociones).

AGRADECIMIENTOS

El proyecto EDECÁN no hubiera visto la luz sin el esfuerzo coordinado de todos los miembros que conforman los equipos investigadores de los cuatro subproyectos, a todos ellos, gracias por el trabajo realizado en la definición y redacción del proyecto. El proyecto EDECÁN lleva activo desde el 31 de diciembre de 2005. En diversas ponencias de estas IV Jornadas en Tecnología del Habla se presentan algunos de los trabajos realizados hasta la fecha.

BIBLIOGRAFÍA

[1] J.F. Allen, D.K. Byron, M. Dzikovska, Towards Conversational Human-Computer Interaction, AI Magazine, 2001

[2] Seneff, S. et al. Galaxy-II: a reference arquitecture for conversational systems development. Proc. ICSLP, 931-934. 1998.

[3] Schwartz, R., Miller, S., Stallard, D., Makhoul, J., Language understanding using hidden understanding models. Proc. of ICSLP. Philadelphia, USA. vol. 2, pp. 997–1000. 1996.

[4] Levin, E., Pieraccini, R., Concept-based spontaneous speech understanding system. Eurospeech'95. Madrid, Spain.

[5] Bonneau-Maynard, H., Lefevre, F. Investigating stochastic speech understanding. En Proc. of ASRU, 260-263. ,2001

[6] Expert Advisory Group on Language Engineering Standards. http://www.spectrum.uni-bielefeld/EAGLES/.

[7] Dialogue Engineering Best Practice Methodology. http://www.disc2.dk. 1999.

[8] Walker, M.A., Litman, D.J., Kamm, C.A., Abella. 1998a. PARADISE: A general framework for evaluating spoken dialogue agents. ACL/EACL 97, pp 271-280.

[9] Beringer, N., Kartal, U., Louka, K., Schiel, F., Turk, U., 2002. PROMISE—a procedure for multimodal interactive system evaluation. In: Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas. pp. 77–80.

[10] Litman, D., Shimei P, 2002. Designing and Evaluating an Adaptive Spoken Dialogue System. User Modeling and User-Adapted Interaction 13, 111-137, 2002.