Demostración del sistema de comprensión de comunicaciones habladas para control de tráfico aéreo del proyecto INVOCA

F. Fernández Martínez, V. Sama Rojo, J. Ferreiros López,
J. Macias-Guarasa, R. De Córdoba, J. M. Montero Martínez,
J. Colas Pasamontes*, E. Campos Palarea*, J. M. Pardo Muñoz
Grupo de Tecnología del Habla, Universidad Politécnica Madrid, Ciudad Universitaria s/n
*Laboratorio de Tecnología Hombre-Computador, Universidad Autónoma Madrid
{efhes, vsama, jfl, macias, cordoba, juancho, pardo}@die.upm.es; jose.colas@ii.uam.es

Resumen: Se presenta un sistema de comprensión de comunicaciones habladas en dos idiomas, castellano e inglés, para el control de tráfico aéreo. Se emplea una arquitectura con dos reconocedores en paralelo más un módulo de detección de idioma. La salida del reconocedor en el idioma elegido pasa al sistema de comprensión basado en reglas dependientes de contexto que extrae los conceptos clave.
Palabras clave: Reconocimiento, multi-idioma, modelado de lenguaje, comprensión, control aéreo.

1 Introducción

El proyecto INVOCA tiene como principal objetivo analizar las posibilidades que ofrece el estado del arte en las tecnologías de reconocimiento de habla para su aplicación a los sistemas de control de tráfico aéreo. Se trata por lo tanto de un proyecto de exploración y evaluación tecnológicas. La funcionalidad del sistema desarrollado es la detección de datos clave en los canales tierra-aire a través de los cuales, tienen lugar las comunicaciones controlador-piloto. Estas están basadas en habla espontánea y responden a una fraseología oficial aplicable a todas las posiciones de control en torre. Pueden producirse tanto en inglés como en castellano por lo que estamos ante un sistema multi-idioma.

2 Descripción del sistema

2.1 Arquitectura

2.1.1 Front-End

El sistema consta de un primer módulo o front-end que convierte la señal acústica en la entrada en el conjunto de vectores de parámetros apropiado. Este front-end esta compuesto a su vez por un detector, cuyo cometido es la detección de voz / no voz a la entrada del sistema, y el parametrizador, que lleva a cabo la parametrización de la información acústica segmentada.

2.1.2 Módulo de reconocimiento

La salida del front-end pasa al módulo de reconocimiento compuesto en este caso por dos reconocedores, uno por cada idioma, inglés y castellano, en que puede haberse realizado la comunicación. Se obtienen las frases reconocidas en el idioma correspondiente para uno y otro reconocedor y se pasan como entradas al módulo de detección de idioma.

2.1.3 Módulo de detección de idioma

El módulo de detección de idioma es el encargado de decidir el idioma al que corresponde la frase que está siendo procesada. La decisión puede tomarse aplicando modelos de lenguaje en base a medidas de perplejidad sobre el resultado de ambos reconocedores. En nuestro caso debido al gran peso dado al modelado de lenguaje durante el proceso de reconocimiento la decisión puede tomarse directamente en base a la diferencia de scores entre ambos reconocedores. Los pesos para el modelado de lenguaje son 9.5 para castellano y 11 para inglés.

2.1.4 Módulo de comprensión

Este módulo es el encargado de extraer la información relevante, los conceptos claves de la tarea a partir de la salida del reconocedor. La comprensión está basada en reglas dependientes de contexto.

Los diccionarios están etiquetados semánticamente en base a las categorías específicas de la tarea. Se cuenta con un diccionario por idioma. A toda la información sin relevancia para la tarea le es asignada una categoría "basura". Se procede de igual forma con las palabras fuera de vocabulario, lo que repercute en una inevitable pérdida de información.

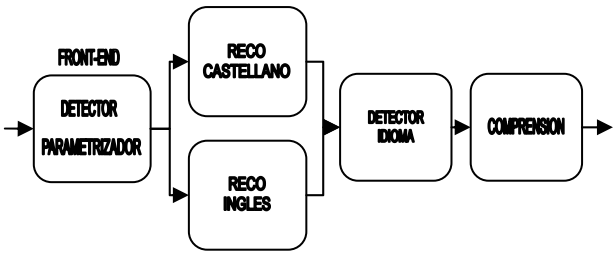

Figura 1: Diagrama de bloques

2.2 Modelos de lenguaje

Se emplean modelos de lenguaje de tipo estocástico basados en bigramas. Se ha trabajado con un corpus de entrenamiento de 3256 frases para inglés y 5091 frases para castellano. El vocabulario resultante se compone de 793 y 1104 palabras respectivamente. Contamos con 36 palabras para las que no disponemos de modelo gramatical en inglés, mientras que en castellano esta cantidad asciende a 86. La no disponibilidad de modelo gramatical para esas palabras es debida a la no aparición de las mismas en entrenamiento. Las perplejidades para cada idioma respectivamente en entrenamiento son 10,4 y 10,5. Sobre un conjunto de test de 453 frases en inglés, y 503 frases en castellano las perplejidades resultantes son 23,2 y 15,2.

3 Resultados obtenidos

Se ofrecen cifras de "*word accuracy"* y *"concept accuracy"*, obtenidos para cada idioma con usuarios reales. También se facilitan datos del % de frases procesadas de forma perfecta, sin errores.

	CASTELL.		INGLÉS	
	%wa	%perf	%wa	%perf
W.A. LIBRE	86'26	33'6	73'26	17'4
C.A. LIBRE	77'73	53'1	47'9	35'5
W.A. GUIADA	96'73	54'3	91'42	19'05
C.A. GUIADA	98'52	91'43	83'94	50

Tabla 1: Resultados obtenidos

En la tabla aparecen resultados para una evaluación libre en la que se procesa todo tipo de frases y para una evaluación guiada en la que sólo se procesan frases que se ajustan a la fraseología entrenada.

Actualmente el proyecto se encuentra en fase de evaluación por lo que los resultados aquí presentados tienen un carácter preliminar. En la actualidad está en proceso la evaluación en la torre de control de Madrid-Barajas. Esta evaluación estará lista en breve y disponible para su presentación en el congreso.

4 Descripción de la demo

4.1 HW necesario

Se contará con dos equipos pc portátiles con sus correspondientes tarjetas de sonido. Se cablearán ambos equipos para llevar el audio de uno a otro.

4.2 Desarrollo de la demo

La demostración consistirá básicamente en la reproducción mediante uno de los equipos de varios ficheros de audio en formato WAV y su correspondiente procesamiento por parte del sistema que estará instalado en el otro equipo. El contenido de dichos ficheros serán grabaciones de las intervenciones de los controladores dentro de comunicaciones reales mantenidas con pilotos.

4.3 Objetivos de la demo

El objetivo principal de la demostración no será otro que el poder observar in situ la respuesta del sistema en condiciones muy similares a las reales con lo que será posible que los asistentes puedan percibir la calidad del mismo. Como proyecto de exploración tecnológica que es, esta demostración puede despertar una gran expectación por ser un buen reflejo del estado del arte de la tecnología de reconocimiento de habla, así como también por sus potenciales prestaciones y aplicaciones en campos tan interesantes como el control de tráfico aéreo.

4.4 Duración de la demo

La duración estimada para la demostración descrita será de aproximadamente 10-15 minutos.