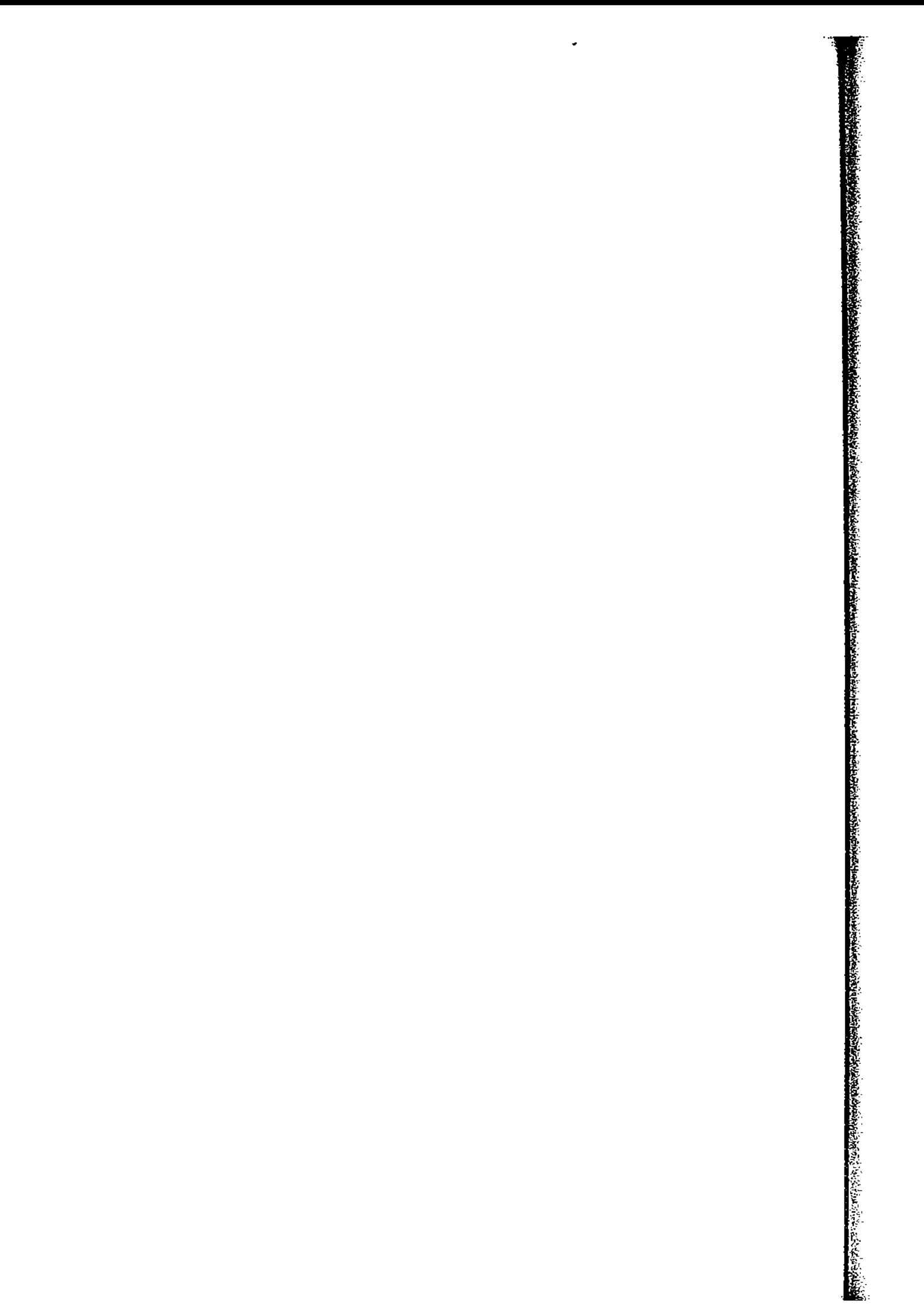


DEMOSTRACIONES



**PRESENTACION DE UNA DEMOSTRACION DE SISTEMAS DE
CONVERSION DE TEXTO A VOZ EN VARIAS LENGUAS.**

*José Manuel Conejo
Bert Van Coile*

Lernout & Hauspie
Speech Products,

Bélgica.

Las demostraciones que proponemos pueden entenderse como una sesión práctica sobre el sistema descrito en la comunicación presentada a la organización del congreso bajo el título "Desarrollo de un conversor de texto a voz en español dentro de una arquitectura multilingüe ". Con estas demostraciones pretendemos ofrecer un panorama del estado de nuestra investigación y desarrollo en la aplicación del sistema a varias lenguas.

Mostraremos conversores con respuesta en tiempo real (Inglés Americano) y el funcionamiento del procesamiento lingüístico de varias lenguas con el sistema DEPES en el la creación de voz artificial a partir de un texto y sin límite de vocabulario.

También se dispondrá de grabaciones de voz sintética.

Merecerá especial atención el status del sistema en español en el momento de la celebración del congreso.

DEMOSTRACION DEL SISTEMA DE CONVERSION DE TEXTO A VOZ PARA CASTELLANO.

Juan Carlos Pérez

Enrique Vidal

Departamento de Sistemas Informáticos y Computación.

Universidad Politécnica de Valencia.

Se propone una demostración interactiva del Sistema de Conversión de Texto a Voz para Castellano que se presenta en la comunicación al congreso de los mismos autores.

Un sistema basado en PC-compatible con un conversor digital-analógico conectado al puerto paralelo del mismo permitirá la introducción interactiva de texto para su lectura así como la lectura de ficheros ASCII.

Se presentará asimismo un prototipo preliminar de la aplicación del sistema a la ayuda al usuario invidente.

**UN RECONOCEDOR DE PALABRAS AISLADAS
EN TIEMPO REAL
PARA UNA ESTACION DE TRABAJO HP-9000
UTILIZANDO MODELOS OCULTOS
DE MARKOV.**

María Teresa Escrig Monferrer.

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia.

En este trabajo se construye un sistema para **Reconocimiento de Palabras Aislada (RPA)** independiente del locutor, para vocabularios reducidos, en tiempo real y sobre una estación de trabajo HP-9000.

A partir de la señal vocal (en nuestro caso, una palabra aislada del vocabulario), se obtiene una representación paramétrica de dicha señal en una etapa denominada **preproceso**, en el cual se realiza un tratamiento mediante **Banco de Filtros**, un cálculo del **cepstrum** y un **etiquetado**, que reducirá la cantidad de información con la que tratar y enfatizará las características más relevantes de la señal.

Esta señal parametrizada servirá para el aprendizaje de la estructura o modelo que representará a esa palabra del vocabulario, utilizando **Modelos Ocultos de Markov (MMO)**. Otras señales parametrizadas podrán ser utilizadas para el reconocimiento, eligiéndose como palabra aquella cuyo modelo ha dado la probabilidad más alta de generación de la muestra a reconocer, según el algoritmo de Forward.

El sistema permite, además, la posibilidad de generación de nuevas tareas, aprendiendo los modelos de cualquier otro vocabulario y la generación de nuevos "codebooks".

1 Este proyecto ha sido parcialmente subvencionado por la Comisión Internacional de Ciencia y Tecnología (CICYII: "Construcción de Sistemas de Reconocimiento del Habla" TIC 448/89.

**SISTEMA DE APRENDIZAJE-RECONOCIMIENTO
EN TIEMPO REAL UTILIZANDO LA TÉCNICA
DEL ALGORITMO ECGI.**

María Isabel Alfonso Galipienso.

Departamento de Sistemas Informáticos y Computación.
Universidad Politécnica de Valencia.
España.

El ECGI (Error Correcting Gramatical Inference) es un algoritmo introducido recientemente como método de aprendizaje, diseñado para lograr una capacidad de abstracción y recoger toda la variabilidad relevante exhibida por la concatenación de sub-estructuras locales de las muestras consideradas, así como sus longitudes (duraciones).

En este caso, las muestras provienen de la señal vocal pronunciada por algún locutor delante de un micrófono; dicha señal, tras sufrir un preproceso para extraer toda la información relevante, es convertida en una cadena de símbolos o etiquetas que constituyen la entrada para el algoritmo de aprendizaje.

El algoritmo ECGI, puede realizar reconocimiento e inferencia de una forma simultánea, lo que se ha aprovechado para reconocer las palabras aprendidas con anterioridad, a medida que el locutor las vaya pronunciando, siempre valiéndose de un micrófono, y obteniendo una respuesta inmediata por pantalla de dicha palabra pronunciada.

INTERFAZ MULTIMODAL: PROYECTO MMI2

SPM

Departament I+D/NLU a ISS
Intelligent Software Solutions, SA
Barcelona

El proyecto MMI2 se desarrolla dentro del marco del Programa Comunitario ESPRIT II (proyecto n. 2474 MMI2: 'Multimodal Interface for Man-Machine Interaction').

El objetivo del proyecto es construir un interfaz adaptable a distintas tipologías de usuarios que disponga de diversos modos integrados de expresión de entrada y salida y con una capacidad suficiente para desarrollar un diálogo robusto y cooperativo entre el usuario y el sistema.

Los modos de expresión que integra el interfaz son los siguientes: gráficos, comandos, gestos y lenguaje natural (inglés, francés y castellano).

La arquitectura del sistema es modular (formada por módulos expertos) y la integración de los diversos módulos y funcionalidades está soportada por un formalismo de representación semántica común a todos los módulos (CMR: Common Meaning Representation) que es usado como vehículo de comunicación interna del sistema.

El prototipo del interfaz funciona con un sistema de bases de conocimiento (SBC) especializado en el análisis y diseño de redes locales y que ha sido desarrollado dentro del mismo proyecto.

El núcleo del interfaz está implementado en BIMprolog y se desarrolla en un entorno de estaciones de trabajo (SUN/3 y SUN/4)¹.

¹ Una presentación general del proyecto MMI2 se encuentra en:

R.Pérez 'MMI2: la construcción de un interfaz multimodal para interacción hombre/máquina'. Artículo de próxima aparición en el boletín de la SEPLN .

Las siguientes comunicaciones presentadas en el VII Congreso de la SEPLN se refieren a trabajos desarrollados por el equipo de ISS SA dentro del proyecto MMI2: Trotzig D. 'Fuerzas ilocucionarias y niveles argumentativos en la generación de diálogo hombre/máquina'.

Pérez R. y Barreras J. 'Un experto semántico para un interfaz multimodal'

DEMOSTRACION DEL MODULO DE LENGUA ESPAÑOLA DEL INTERFAZ MULTIMODAL MMI²

SPM

Departament I+D/NLU a ISS
Intelligent Software Solutions, SA
Barcelona

SPM es el módulo de lengua española para la comunicación usuario-sistema en lengua española del interfaz MMI² diseñado y desarrollado en el departamento de I+D/NLU de ISS.

El interfaz MMI² se construye dentro del proyecto MMI²: 'A Multi Modal Interface for Man-Machine Interaction with Knowledge Based Systems' en el marco del Programa ESPRIT de la Comisión de las Comunidades Europeas, en el que ISS participa dentro de un consorcio de empresas y centros públicos de investigación europeos.

El objetivo del proyecto es la construcción de un interfaz hombre/máquina para diferentes tipos de usuarios que integre diversos modos de comunicación: lenguajes naturales (inglés, francés y español), comandos, diversos modos gráficos y gestual.

El proyecto pone especial énfasis en la integración multimodal y en una arquitectura flexible que permita su portabilidad a otras aplicaciones.

En la actualidad se ha puesto a punto un prototipo integrado de demostración del interfaz MMI² que será presentado en la 'ESPRIT Technical Week' de Bruselas (noviembre de 1991). El demostrador ha sido conectado a un sistema experto en el diseño de redes locales. SPM se halla integrado en el demostrador MMI² como su modo de comunicación en lengua española.

El diseño del SPM combina aspectos y principios metodológicos generales desde diversas perspectivas

- la ingeniería de software.-
la integración funcional modular,
la reusabilidad modular,
la eficiencia en el rendimiento modular.
- la ingeniería de interfaces.-
la integración multimodal,
la adaptabilidad al usuario
la cooperatividad intermodular.
- la ingeniería de la comunicación lingüística.-
la robustez,
la cooperatividad,
la adaptabilidad de dominios y sublenguajes,
- la ingeniería lingüística.-
la arquitectura lingüística modular: morfología, sintaxis,
semántica y pragmática.

el cubrimiento lingüístico fenoménico,
la precisión de la representación del significado,
la composicionalidad como principio operativo del análisis.

La demostración destacará aspectos del diseño relacionados con los dos últimos apartados, en especial, las facilidades en la gestión flexible de la información lingüística para la introducción de conceptos de dominio y sublenguaje, y los mecanismos de interacción del tratamiento de la cadena lingüística con el conocimiento top-down sobre el dominio y el sublenguaje para reducir la ambigüedad de análisis, evitar el fallo y reparar el error.

La demostración presentará la riqueza del análisis lingüístico y el cubrimiento de fenómenos lingüísticos, así como señalará otras funcionalidades como por ejemplo, la cooperatividad en la resolución de palabras desconocidas, los mecanismos de decisión ante fenómenos de ambigüedad. El auditorio apreciará también la cualidad y eficiencia del procesamiento lingüístico.

Se demostrarán los módulos que componen el SPM:

MORFEO

Lleva a cabo el análisis morfológico y desarrolla un proceso de lematización del input.

Consta de tres submódulos:

1º Morfografía: un inicial tratamiento de signos ortográficos.

2º Analizador morfológico: análisis de la información morfológica dinámica, la categoría morfológica y la lematización.

3º Filtro métrico: aplicador de la métrica del sublenguaje y los mecanismos de constitución de ese conocimiento.

LEX

El proceso de constitución del léxico de 'runtime' donde interviene un diccionario virtual estándar, el Experto Semántico del interfaz y otros ficheros de información menores. La estructura de la información léxica. La combinación de información léxica general estándar y de información léxica top-down del dominio.

SP

El parser de lengua española. Su proceso de análisis ascendente siguiendo un mecanismo 'left-corner'.

La gramática y su cubrimiento lingüístico: concordancia, adscripción, control, relativas, coordinación argumental, pasiva, pasiva refleja, comparativos, etc.

La incrementalidad y principio de composicionalidad del análisis.

El análisis de la información ilocucionaria del significado.

LOGIC

La traducción de descripciones funcionales al formalismo de base lógica CMR (Common Meaning Representation) que es el formalismo común de representación del significado de la comunicación para todos los modos del interfaz en cuanto a la consecución del objetivo de la gestión integrada del diálogo multimodal.

Los mecanismos del paso de la estructura de dependencias a la CMR: cálculo del alcance de los cuantificadores y reificación de predicados.

GENIUS

El módulo de generación del SPM.

Los mecanismos múltiples de generación de comunicaciones en español.

EL PROCESO DE RECONOCIMIENTO DEL HABLA CON UN AMPLIO LEXICO

Anne DEMEDTS

Dragon Systems, Inc en Boston

En 1989 Dragon Systems, Inc. presentó la primera versión del DragonDictate para el inglés americano en Speech Tech en Nueva York : se trata de un sistema de reconocimiento del habla que se adapta a la fonética particular del usuario y que reconoce en tiempo real un amplio léxico de 30.000 unidades en una dicción discreta. Actualmente se están realizando las versiones española, francesa, alemana, italiana y neerlandesa, cuya finalización está prevista para 1992.

Según esta tecnología, primero hay que crear modelos acústicos de todos los fonemas de una lengua determinada, considerados en la variedad de todos sus contextos posibles. A partir de ahí se elaboran Modelos de Markov Escondidos (H.M.M. : Hidden Markov Models) cuyos parámetros pueden ser re-evaluados a base de una información fonética mínima. Se ha alcanzado ya la fase de comparación entre prototipos en las distintas lenguas mencionadas : operan con un vocabulario reducido de unas siete mil palabras logrando un rendimiento comparable al de

El léxico básico del DragonDictate está compuesto de 30.000 unidades fonéticas, de las cuales 5.000 son escogidas por el usuario, que las añade al núcleo de las 25.000 unidades más frecuentes de una lengua particular. Así~, por ejemplo, para el inglés americano se ha recurrido a un análisis estadístico de materiales facilitados por la agencia UPI. El indicio de frecuencia acompaña a la ortografía de cada unidad así como un modelo acústico que el ordenador construye a partir del análisis de la voz tal como fue pronunciada por un hablante nativo. Estos datos estadísticos son actualizados constantemente de acuerdo con el idiolecto del usuario. En el proceso de reconocimiento intervienen tres componentes básicos, que suponen una aproximación gradual entre un conjunto de probabilidades y la palabra dicha por el usuario.

Son un algoritmo de verificación rápida, otro algoritmo de programación dinámica y el ya referido modelo lingüístico estadístico.

EL LEXICO BASICO

En cualquier momento DragonDictate dispone de un conjunto de 30.000 palabras caracterizadas en cuadros de 20 milisegundos por medio de 8 parámetros acústicos. Esto no implica que un hablante de referencia haya tenido que decir las todas ni para el usuario que haya que hacer lo mismo a fin de personalizar el léxico porque un vocablo se compone de cierto número de 'fonemas en contexto' (PIC : Phonemes In Context) que no le son privativos. A su vez un fonema siempre consta de, hasta 6 diferentes, elementos fonéticos (PEL : Phonetic Element) que se encuentran compartidos por varios alófonos del mismo. De esta manera se entablan relaciones, a menudo muy complejas, entre diferentes unidades. También explica que DragonDictate 'aprende' a medida que el usuario lo utilice, gracias a esta extrapolación de datos.

A esta información fonética básica se añade otra de tipo durativo relacionada con tanto la ubicación del acento como la estructura de la palabra. Considérese, por ejemplo, que en español la vocal acentuada es relativamente larga en palabras agudas que no terminen en 'n' o 'l' ("papá"), mientras la vocal inacentuada es generalmente breve. Si bien es cierto que el papel de estos factores varía, las mismas herramientas informáticas son válidas para cada lengua.

EL PROCESO DE RECONOCIMIENTO

Tampoco los componentes integrantes del mecanismo de reconocimiento dependen fundamentalmente de la lengua utilizada.

El algoritmo de verificación rápida efectúa una primera selección dentro del conjunto léxico comparando la secuencia inicial de la unidad detectada con la de unos cientos o, aun, miles de grupos en los que ha reunido previamente palabras que empiezan de la misma forma. Así el número de palabras eventualmente dichas se reduce a unos doscientas.

El algoritmo de programación dinámica hace uso de H.M.M. estableciendo una compleja red de probabilidades entre diversos PELs acompañados por la expresión en milisegundos de su duración calculada en base al entorno lingüístico.

De momento la función del modelo lingüístico se limita a la estimación de la probabilidad de una palabra, teniendo en cuenta la palabra anterior. Se basa en un cálculo de frecuencias, que podrá incluir también frecuencias de pares de palabras.

EL ESPAÑOL

En cuanto a reconocimiento bastan para la caracterización fonética del español 11 vocales y diptongos, y 22 consonantes. Inicialmente se tiene prevista la inclusión de 3 tipos de acentuación por analogía con, entre otros, el neerlandés y el inglés. Para facilitar la comparación entre los sistemas en idiomas diferentes, se han tomado como punto de partida los capítulos 1, 2, 3, 7 y 8 del texto de A. de Saint-Exupéry "Le Petit Prince" y sus traducciones a las lenguas respectivas. A estos se les van añadiendo textos auténticos en cada una de las lenguas.

Pruebas con el DragonDictate para el inglés americano pusieron de relieve que tras una adaptación de aproximadamente 2.000 palabras, entre el 85% y el 90% de las palabras son reconocidas correctamente. En cuanto a los errores se distinguen los tipos siguientes :

- "opciones" (O) : la palabra correcta figura en una lista de hasta 8 opciones alternativas que aparece en la pantalla de modo que el usuario puede corregir el error pulsando una tecla
- "nueva palabra" : todavía no existe un modelo acústico para la palabra puesto que no forma parte del léxico básico del sistema. Por tanto, el usuario habrá de escribir toda la palabra : a continuación se elabora un modelo acústico que el sistema podrá relacionar con la grafía en cuestión.
- "error" (E) la palabra no figura entre las opciones y el usuario tiene que introducir uno o más caracteres antes de que sea reconocida correctamente.

Por lo que respecta al prototipo español se han llevado a cabo varios experimentos.

En una primera prueba intervenían tres locutores : el hablante de referencia que habla el español peninsular (1), otro hablante nativo procedente de Venezuela (2) y, finalmente, una

persona cuya lengua materna no es el español (3). Se indica el sexo de los hablantes entre paréntesis. En el desarrollo de la prueba se sucedían tres etapas : lectura del primer capítulo de "El Principito" - fase de iniciación -, lectura de los capítulos 2 y 3 - fase de adaptación -, lectura de los capítulos 7 y 8 - fase de evaluación del rendimiento -.

Como se desprende del esquema, aunque inicialmente el rendimiento es mejor para el hablante de referencia el sistema se adapta al acento peculiar del usuario. Además, pruebas en otras lenguas demuestran que incluso la divergencia de sexo entre el hablante de referencia y otros usuarios no plantea mayores dificultades. (Las cifras indican porcentajes)

	cap. 1			cap. 2, 3			cap. 7, 8		
	O	N	E	O	N	E	O	N	E
1. AD (f)	83	12	5	86	10	4	88	7	5
2. KK (f)	76	9	15	85	7	8	86	6	8
3. KH (f)	76	12	12	84	7	9	87	6	7
POR MEDIO	78	11	11	85	8	9	87	6	7

Durante la fase inicial el hablante de referencia obtiene, claro está, el mejor rendimiento ; sin embargo, basta con dictar unas 2000 palabras para que el sistema se haya adaptado a la fonética de cada usuario. No hace falta que la pronunciación sea 'correcta', tan solo habrá de ser consistente.

En el caso de los dos usuarios otros que el hablante de referencia, el porcentaje de verdaderos 'errores' - en que la unidad dictada ni siquiera figuraba en la lista de opciones - es el doble del porcentaje análogo para el hablante de referencia. Esto sugiere que existan cualidades propias de los modelos de verificación rápida fuera del alcance de la adaptación.

Tan solo para la versión inglesa del DragonDictate disponemos ya de la versión con un amplio vocabulario de reconocimiento (30.000 unidades). La comparación de los resultados obtenidos, por una parte, con un prototipo reducido y, por otra, con la versión de léxico amplio pone de relieve que un aumento considerable del vocabulario no afecta al rendimiento.

SISTEMA DE RECONOCIMIENTO DE PALABRAS AISLADAS EN TIEMPO REAL PARA UN COMPUTADOR PERSONAL

Juan Antonio Puchol García

DSIC - UPV Valencia

En este proyecto se presenta un Sistema de Reconocimiento de Palabras Aisladas en tiempo real, usando una placa específica de procesamiento de señal, llamada DSP-16, que ha permitido realizar toda la fase de preproceso de la señal en tiempo real. El preproceso consta de las siguientes fases:

- Preproceso Analógico.
- FFT (Fast Fourier Transform) empleando una ventana Hamming.
- Bancos de Filtros (Segun escala de Mel).
- Detección de bordes (usando un buffer cíclico).
- Coeficientes Cepstrales.

Para la fase de reconocimiento se ha empleado un algoritmo de **alineamiento temporal no lineal**, derivado por programación dinámica. El empleo de ventanas de ajuste, junto con una representación compacta de las palabras, ha posibilitado reducir el coste temporal de dicho algoritmo, contribuyendo a que la fase de reconocimiento se pueda realizar también en tiempo real.

Para reconocer las palabras se emplea un diccionario que debe ser adquirido previamente a la fase de reconocimiento. Se ha empleado como **Regla de Decisión**, para saber qué palabra es la que se ha pronunciado, el clasificador del vecino mas próximo, usando como funcion de disimilitud (o distancia) el resultado que da el alineamiento temporal no lineal.

Toda la aplicación se apoya en un entorno desarrollado enteramente en Turbo Pascal 6.0 de tipo *Windows*.

