DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TESIS DOCTORAL

ESTRATEGIAS DE INCORPORACIÓN DE CONOCIMIENTO SINTÁCTICO Y SEMÁNTICO EN SISTEMAS DE COMPRENSIÓN DE HABLA CONTINUA EN CASTELLANO

José Colás Pasamontes

Ingeniero de Telecomunicación

Director de la Tesis

Dr. Ingeniero José Manuel Pardo Muñoz

1.999

Tesis Doctoral: **ESTRATEGIAS DE INCORPORACIÓN DE CONOCIMIENTO SINTÁCTICO Y SEMÁNTICO EN SISTEMAS DE COMPRENSIÓN DE HABLA CONTINUA EN CASTELLANO**

Autor: **JOSÉ COLÁS PASAMONTES**

Director: **Dr. INGENIERO JOSÉ MANUEL PARDO MUÑOZ**

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores:

Presidente: JAVIER FERREIROS

Vocales: NATIVIDAD PRIETO
LUIS A. HERNANDEZ
ANTONIO RUBIO

Secretario: JOSE CARLOS GONZALEZ

acuerda otorgarle la calificación de Sobresaliente cum Laude (5 votos)

Madrid, a 1 de julio de 1.999

RESUMEN

Este trabajo de tesis doctoral se ha planteado analizar la problemática del proceso de **comprensión de habla**, no sólo desde un punto de vista científico sino también técnico, concluyendo con el **diseño, implementación y evaluación de un Sistema de Comprensión de Habla en castellano**.

En el marco de esta tesis se ha realizado una **revisión de distintas soluciones que han sido propuestas por diversos grupos de investigación internacionales para resolver el problema de la comprensión de habla**.

Se ha definido una **arquitectura no integrada novedosa para la comprensión del habla en castellano,** es decir, teniendo en cuenta características del castellano como lengua natural que no aparecen o aparecen menos acentuadas en otras lenguas. Esta arquitectura pretende ser la base de futuros trabajos en esta línea en el Grupo de Tecnología del Habla, de la Universidad Politécnica de Madrid. Con el objetivo de comprender habla perteneciente a un **dominio semántico restringido** (limitado por los conceptos que existen en el dominio de una aplicación concreta), esta arquitectura tiene como características principales:

- La **robustez**, es decir, la posibilidad de procesar frases que contienen errores (inserciones, borrados o sustituciones de palabras) producidos por el módulo de decodificación acústica (sistema de reconocimiento de habla continua), o que tienen rasgos de agramaticalidad producidos por la propia naturaleza del lenguaje hablado, o con problemas de cobertura a nivel léxico, sintáctico o semántico.

- La **modularidad**, que permite seguir mejorando sin necesidad de rediseñar e implementar el sistema completo.

- La **flexibilidad**, con el fin de independizar la arquitectura de una aplicación concreta dentro, lógicamente, de ciertas restricciones, impuestas por la naturaleza de los sistemas de información o sistemas de control automáticos.

- La **potencia**, definida como la posibilidad de procesar frases de un cierto nivel de complejidad lingüística.

Los diferentes módulos incorporan conocimiento lingüístico de distinta naturaleza, lo que ha permitido **estudiar la interacción de distintas fuentes de conocimiento lingüístico y un modo eficaz de integrarlas**, en el proceso de comprensión.

Se ha utilizado **información semántica en forma de rasgos**, que completan la ya modelada por las categorías semánticas del diccionario, a **gramáticas contextuales** simplificadas (definiendo **lenguajes específicos de reglas y algoritmos de análisis o ejecución de estas reglas**), que en forma de reglas solucionan principalmente problemas de ambigüedad semántica y elipsis, y una **gramática semántica de contexto libre** (utilizando el algoritmo de Earley con capacidad para procesar frases con ambigüedad) que pretende, basándose en una clasificación o taxonomía de los

conceptos del dominio que reduce en gran medida las reglas necesarias, obtener la información estructural de las mismas que ayuda al procesamiento de frases de una cierta complejidad manteniendo el proceso de **traducción a SQL, necesario en sistemas de información con acceso a bases de datos, dentro de unos límites de sencillez sorprendentes**, mediante el uso de **plantillas semánticas**.

Con el fin de evaluar el comportamiento del módulo de decodificación acústica se ha implementado un sistema de reconocimiento de habla continua modular, con capacidad para integrar conocimiento gramatical en base a cualquier gramática probabilística de tipo N-gram, de naturaleza morfo-sintáctica o semántica. Se han evaluado distintas gramáticas guiando el proceso de decodificación acústica. Con el fin de mantener la eficacia del este módulo a pesar de la incorporación de gramática en el proceso, **se ha estudiado con profundidad un mecanismo de reducción del espacio de búsqueda ampliamente utilizado conocido como "recorte de caminos" o "recorte del haz" (*beam-search*)**, presentando un método que se ha desarrollado en esta tesis que permite la determinación del *umbral de recorte* basado en la probabilidad (o distancia) del mejor estado del espacio de búsqueda en cada trama (estadístico) de antemano, utilizando los datos de entrenamiento y conociendo la influencia que tendrá dicho umbral en el proceso de reconocimiento. Además, se han evaluado dos variantes ya conocidas: el uso de *uno o dos umbrales de recorte,* uno basado en la probabilidad (o distancia) del mejor de los últimos estados de cada modelo en cada trama (estadístico del último estado) y otro en la probabilidad (o distancia) del mejor del resto de los estados distintos del último en cada trama (estadístico del resto de los estados), y se han aportado nuevas conclusiones al respecto. Todo ello ha permitido profundizar en el funcionamiento de esta técnica ya conocida pero no tan estudiada. Además, el decodificador acústico ha sido modificado para permitir la generación de varias hipótesis (frases) de salida (las N mejores), y **se ha estudiado la relación entre el valor de N (número de caminos o hipótesis) y la calidad del sistema de reconocimiento (mejora de la tasa de acierto de palabras o reducción del error del sistema)**, para aplicaciones como la que ha sido objeto en esta tesis. **Se ha comprobado que con un número de hipótesis reducido (N muy pequeño) se consigue que el módulo acústico se recupere de muchos errores que afectarían al proceso de comprensión de la frase hablada reconocida.**

ABSTRACT

This Ph.D. thesis work is aimed at analyzing the problems when facing **automatic speech understanding**, from both scientific and technical points of view, concluding with the **design, implementation and evaluation of a Castilian Spanish Speech understanding system.**

In this thesis, **some of the alternatives, that have been proposed by international research groups in order to solve the speech understanding problem, have been reviewed.**

A **novel non integrated architecture for speech understanding in Spanish** has been defined, taking into account the specific characteristics of Spanish as a natural language, not found or rarely found in other languages. This architecture intends to be the baseline of future work in this topic in the Speech Technology Group, in the Universidad Politécnica de Madrid.

To achieve the objective of understanding speech in **limited semantic domains** (limited by the concepts used in the domain of a specific application) this architecture has been designed with the following main characteristics:

- **Robustness**, that is, the possibility of processing sentences with errors (word insertions, deletions or substitutions) produced by the acoustic decoding module (a continuous speech recognition system); or non-grammatical constructions, due to the inherent characteristics of spoken language; or problems in lexical, syntactic o semantic coverage.

- **Modularity**, that permits improving the system without redesigning or implementing the whole system.

- **Flexibility**, in order to have an application-independent architecture, obviously under certain restrictions, imposed by the characteristics of both automatic information and control systems.

- **Power**, defined as the possibility of processing sentences with a certain degree of linguistic complexity.

Those modules incorporate linguistic knowledge of different kinds, and this has allowed us to **study the interaction of different linguistic knowledge sources and an efficient way of integrating them** in the understanding process.

Features to represent the semantic information have been used, completing the one already modelled by the dictionary semantic categories; simplified **contextual grammars** (defining **specific rules languages, and rule analysis or execution algorithms**), which mainly solve some of the semantic ambiguity and ellipsis problems; and a **semantic context free grammar** (using the Earley algorithm with its possibility of processing ambiguous sentences). The latter intends to obtain the structural information of the sentence, using a taxonomy of the domain concepts that heavily reduces the number of needed rules. Moreover, it helps the processing of complex

sentences, while keeping the SQL translation process surprisingly simple, by using **semantic templates**. This translation process is **needed in information systems accessing databases**.

In order to evaluate the acoustic decoder module behaviour, a modular continuous speech recognition system has been implemented. It is able to integrated grammatical knowledge based on any stochastic morpho-syntactic or semantic N-gram. To keep the efficiency of this module, even when the grammar information is used, **a search space reduction mechanism (beam-search) has been deeply studied**. A new method developed in this Thesis allows to analyse and to determine, in advance, a *pruning threshold* based on the probability (or distance) of the best state in the search space for every frame (stochastic), making use of training data and knowing the impact this threshold will have in the recognition process. Besides, two well-known variants have been evaluated: the use of *one or two pruning thresholds*, one based on the probability (distance) of the best last states for every model in every frame (stochastic parameter of the last state) and the other one based on the probability of the best of the rest of the states in every frame (stochastic parameter of the rest of the states). New conclusions have been drawn from this study. All this allowed us to deepen in this well known but not so well understood technique. Moreover the acoustic decoder has been modified to allow the generation of several output hypothesis (N-best sentences), and **the relationship between the value N (number of paths or hypothesis) and the speech recognition system performance (improvement of the word error rate)**, for applications such as the one aimed in this Thesis. **We have checked that with a small number of hypothesis (very low N), the acoustic module is able to recover from a lot of errors that would severely affect the understanding process of the recognised spoken sentence.**

ESTRATEGIAS DE INCORPORACIÓN DE CONOCIMIENTO SINTÁCTICO Y SEMÁNTICO EN SISTEMAS DE COMPRENSIÓN DE HABLA CONTINUA EN CASTELLANO

ÍNDICE

CAPÍTULO 6. EL MÓDULO DE COMPRENSIÓN

CAPÍTULO 7. CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS

APÉNDICES