Speech into Sign Language Statistical Translation System for Deaf People

B. Gallo, R. San-Segundo, J.M. Lucas, R. Barra, L.F. D'Haro, F. Fernández

***Abstract*— This paper presents a set of experiments used to develop a statistical system from translating speech to sign language for deaf people. This system is composed of an Automatic Speech Recognition (ASR) system, followed by a statistical translation module and an animated agent that represents the different signs. Two different approaches have been used to perform the translations: a phrase-based system and a finite state transducer. For the evaluation, the followings figures have been considered: WER (Word Error Rate), BLEU and NIST. The paper presents translation results of reference sentences and sentences from the Automatic Speech Recognizer. Also three different configurations have been evaluated for the Speech Recognizer. The best results were obtained with the finite state transducer, with a word error rate of 28.21% for the reference text, and 29.27% using the ASR output.**

***Keywords*— Statistical Machine Translation, Sign Language, phrase, Finite State Transducer, Language Model, Translation Model, alignment, word error rate.**

I. INTRODUCCIÓN

Con la realización de este trabajo se pretende el desarrollo y evaluación de una Plataforma de Traducción capaz de transformar, en base a un conjunto de modelos probabilísticos, frases de castellano a Lengua de Signos Española (LSE). Su importancia radica en la necesidad de una herramienta que permita una traducción rápida y precisa entre lenguas. En relación directa con la traducción de castellano a Lengua de Signos Española, se encuentra el campo de aplicación relacionado con personas sordas, donde el coste de un intérprete signante (que conoce la Lengua de Signos) es muy elevado a la vez que se intenta ayudar a las personas sordas prelocutivas (aquellas que se quedaron sordas antes de poder hablar), ya que su capacidad de comprensión del español es inferior a la de los oyentes.

Este artículo se centra en la Traducción (Automática) Estadística o "SMT" ("Statistical Machine Translation"), que es un paradigma de traducción automática donde se generan traducciones en base a modelos estadísticos y de teoría de la información cuyos parámetros se obtienen del análisis de corpus de textos bilingües (documentos que constituyen la base de datos con pares de frases castellano-LSE). Para la realización de los experimentos se dispuso de unos textos en castellano y en LSE con frases típicas que un funcionario pronuncia en el servicio de solicitud o renovación del DNI o el pasaporte.

Este trabajo ha sido posible gracias a la financiación de los siguientes proyectos: EDECAN (MEC Ref: TIN2005-08660-C04), ROBONAUTA (MEC Ref: DPI2007-66846-C02-02) y ANETO (UPM-DGUI-CAM. Ref: CCG07-UPM/TIC-1823).

B. Gallo, R. San-Segundo, J.M. Lucas, R. Barra, L.F. D'Haro, F. Fernández trabajan en el Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la Universidad Politécnica de Madrid, Madrid SPAIN, lapiz@die.upm.es.

II. ARQUITECTURA DEL SISTEMA

El sistema completo está formado por tres módulos: el del reconocedor de voz, el módulo de traducción estadística y finalmente, la representación de los signos mediante un avatar. Se muestra a continuación el diagrama del sistema:

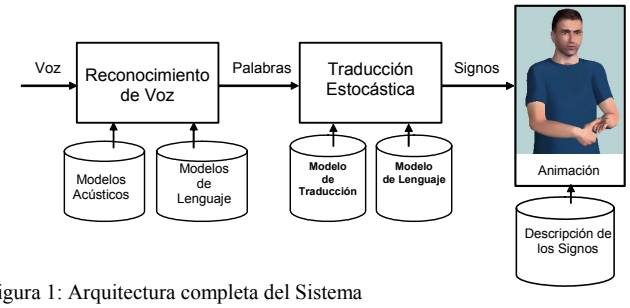

Figura 1: Arquitectura completa del Sistema

El reconocedor del habla realiza la conversión del lenguaje natural (habla continua) a una secuencia de palabras basándose en unos Modelos de Lenguaje y Acústicos de los que se dispone. El módulo de traducción estadística consiste en un algoritmo de búsqueda dinámica que utiliza un modelo estadístico para obtener la mejor secuencia de signos resultado de la traducción de una secuencia de palabras obtenidas del reconocedor de voz. Este modelo integra información de dos tipos de probabilidades: la probabilidad de traducción, que recoge información sobre qué palabras se traducen por qué signos y la probabilidad de secuencia de signos, que aporta información sobre qué secuencias de signos son más probables en la LSE. El último módulo corresponde al avatar en 3D, que se encarga de la representación de los signos provenientes de la Traducción Estadística. El avatar utilizado es "VGuido" del proyecto eSIGN [1].

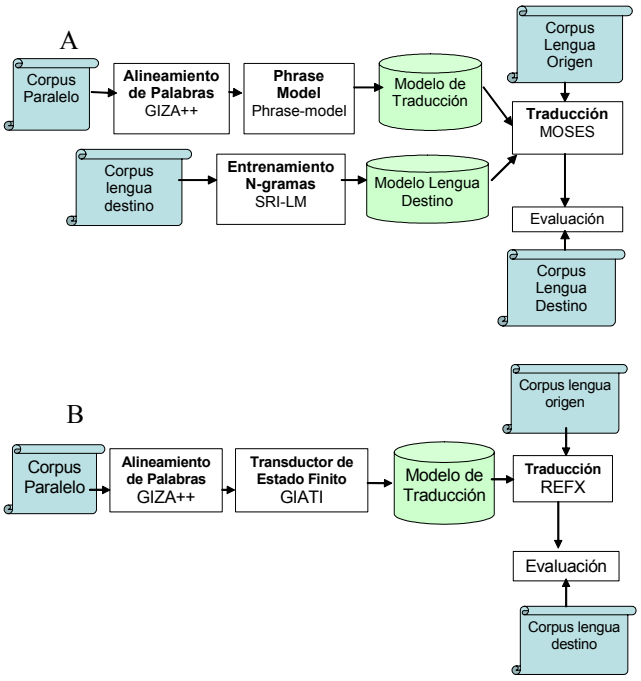

Fig.2: (A) Arquitectura completa de la Traducción basada en Subsecuencias de palabras. (B) Arquitectura completa de la Traducción basada en Transductores de Estados Finitos. Todas las herramientas utilizadas para la traducción son de software libre.

III. Traducción estadística basada en Modelos de Subsecuencias de Palabras

La traducción estadística basada en modelos de subsecuencias (o subfrases) consiste en la obtención de un Modelo de Traducción a partir del alineamiento y extracción de subsecuencias utilizando un corpus paralelo, y la generación de un modelo de lenguaje de la lengua destino. Después se obtiene la secuencia de signos dada una frase de entrada, que luego se evalúa para calcular los aciertos y fallos en la traducción. La arquitectura completa se muestra en la Fig.2.A.

A. Generación de modelos

En primer lugar debe crearse el Modelo de Lenguaje de la lengua destino y el Modelo de Traducción (a partir de un corpus paralelo tanto en lengua origen (LO) como destino (LD)). El problema de la traducción se centra en conocer la probabilidad p(d|o) de que una cadena o de LO genere una cadena d en LD. Estas probabilidades se calculan utilizando técnicas de estimación de parámetros a partir del corpus paralelo. Aplicando el Teorema de Bayes a p(d|o) esta probabilidad se representa como el producto p(o|d)·p(d), donde el Modelo de Traducción p(o|d) es la probabilidad de que la cadena origen se traduzca por la cadena destino, y el Modelo de Lenguaje p(d) es la probabilidad de ver aquella cadena origen. Matemáticamente, encontrar la mejor traducción $\tilde{o}$ se calcula como:

$$\tilde{o} = \arg\max_{o \in o^*} p(d/o) = \arg\max_{o \in o^*} p(o/d)\cdot p(d) \quad (1)$$

Para la creación del Modelo de Lenguaje, se utiliza la herramienta SRILM [2], que realiza la estimación de los modelos de lenguaje tipo N-grama (en la que la probabilidad de una palabra depende de las N anteriores), a partir del corpus de entrenamiento. La generación de los Modelos de Traducción se hace mediante una traducción basada en subfrases. Para esto la herramienta utilizada es el GIZA++ (que es una implementación de los modelos IBM de traducción [3]), un sistema de traducción estadística automática capaz de entrenar estos modelos para cualquier par de lenguas [4]. Para esto se necesita una colección de textos traducidos, que será el corpus paralelo. Los pasos para la generación de los modelos son:

1. Obtención del alineamiento entre palabras: a partir de los dos textos en castellano y LSE se identifican qué palabras de uno se alinean con los signos de LSE. El alineamiento se calcula en ambos sentidos: palabras-signos y signos-palabras.
2. Cálculo de una tabla de traducción léxica: se calcula a partir del alineamiento, obteniendo los valores de w(d|o) y su inversa w(o|d) para todas las palabras, es decir, las probabilidades de traducción para todos los pares de palabras.
3. Extracción de subsecuencias de palabras: se recopilan todos los pares de subsecuencias que sean consistentes con el alineamiento.
4. Cálculo de las probabilidades de traducción de cada subsecuencia: se calculan las probabilidades de traducción para los pares de subfrases en los dos sentidos: subfrase en castellano- signo en LSE y signo en LSE – subfrase en castellano.

B. Ajuste y traducción

Para realizar el proceso de traducción se combinan los modelos generados en la fase anterior de entrenamiento mediante una combinación lineal de probabilidades cuyos pesos se deben ajustar. Este proceso consiste en probar el traductor Moses [4] con un conjunto de frases (conjunto de validación) y, conociendo la traducción correcta, evaluar las salidas del traductor automático en función de los diferentes valores asignados a los pesos. Finalmente, y utilizando un nuevo conjunto de frases (conjunto de test) se evalúa el sistema. Moses es un sistema de traducción automática estadística basado en subsecuencias de palabras, que implementa un algoritmo de búsqueda para obtener, a partir de una frase de entrada, la secuencia de signos que con mayor probabilidad corresponde a su traducción.

IV. Traducción estadística basada en Transductores de Estados Finitos

Los transductores de Estados Finitos ("FST") parten de un corpus de entrenamiento y, usando métodos de alineamiento basados en GIZA++, generan un conjunto de cadenas a partir de las cuales se puede inferir una gramática racional. Esta gramática se convierte en un FST caracterizado por su topología y distribuciones de probabilidad, características aprendidas con el programa GIATI [5]. En la Fig.2.B se muestra la arquitectura de esta solución. Los pasos de esta estrategia son los que se explican a continuación.

A. Alineamiento con GIZA++ (similar al apartado III.A.1.)

B. Transformación de pares de entrenamiento a frases

Se construye ahora un *corpus extendido* a partir de cada uno de los pares de subsecuencias de entrenamiento y sus correspondientes alineamientos obtenidos con GIZA++: se asignarán por tanto palabras de LO a su correspondiente palabra en LD gracias a su alineamiento. Se muestra a continuación un ejemplo de pares castellano / LSE y su alineamiento:

el denei es obligatorio desde los catorce años # DNI(2) SE-LLAMA(3) OBLIGATORIO(4) DESDE(5) CATORCE(7) PLURAL(6) AÑO(8) EDAD(8)

el denei es el documento oficial # DNI(2) SE-LLAMA(3) DOCUMENTO(5) OFICIAL(6)

A continuación se forman las *palabras extendidas* ("extended words", unión de palabras y signos alineados), que representan la traducción propuesta. En este ejemplo:

(el, λ) (denei, DNI) (es, SE-LLAMA) (obligatorio, OBLIGATORIO) (desde, DESDE) (los, PLURAL), (catorce, CATORCE) (años, AÑO EDAD)

(el, λ) (denei, DNI) (es, SE-LLAMA) (el, λ) (documento, DOCUMENTO) (oficial, OFICIAL)

C. Inferencia de un Gramática Estocástica y posteriormente de un Traductor de Estados Finitos

Se obtiene un FST a partir de las frases con las palabras extendidas. Las probabilidades de saltos entre nodos de un FST se computan por las cuentas correspondientes en el conjunto de entrenamiento de palabras extendidas. Se ilustra este proceso en la siguiente figura, donde los nodos grises indican que la frase de salida puede terminar en ese punto:

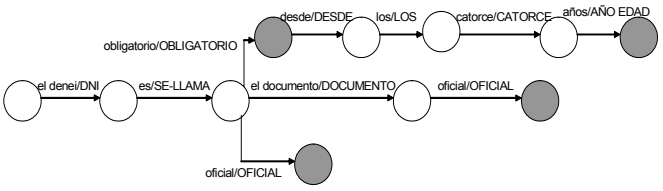

Figura 3: Transductor de estado finito inferido a partir del bigrama del ejemplo anterior

V. EVALUACIÓN DE LAS ALTERNATIVAS TECNOLÓGICAS

A. Medidas de evaluación

Para evaluar la calidad de la traducción, se calculan las medidas de evaluación que se explican a continuación. WER ("Word Error Rate", porción de palabras con error) se calcula como: WER = (S+B+I) / N, donde S es el número de sustituciones, B es el número de borrados, I es el número de inserciones y N es el número de palabras que tiene la salida de la traducción en relación con la frase de referencia. BLEU [6] es un método que evalúa la calidad de las traducciones con respecto a otras de referencia a nivel de frase. Halla la precisión en N entre la traducción del sistema y la de referencia. NIST [7] es similar al BLEU pero utiliza una media aritmética para reducir el impacto de bajas concurrencias para órdenes altos de N-gramas y considera la calidad de la información que proporciona un N-grama particular.

B. Base de Datos

La base de datos utilizada para los experimentos consiste en un corpus paralelo que contiene 414 frases típicas de un contexto restringido: aquellas que diría un funcionario cuando asiste a gente que quiere renovar el pasaporte y/o el Documento Nacional de Identidad, o información relacionada. En este contexto concreto, un sistema de traducción de voz a LSE es muy útil puesto que la mayoría de estos empleados no conocen este lenguaje y tienen dificultades a la hora de interactuar con personas sordas.

El conjunto de frases se dividió aleatoriamente en tres grupos: entrenamiento (conteniendo aproximadamente el 70% de las frases), evaluación (con el 15% de las frases) y test (15% de frases). Esta concentración se hace de forma arbitraria. Se muestra a continuación un resumen de la base de datos:

TABLA I
ESTADÍSTICAS DE LA BASE DE DATOS

		Castellano	**LSE**
Total	**Pares de Frases**	**414**	
	Nº de palabras/glosas	**4847**	**4564**
Entrena-miento	Pares de Frases	314	
	Nº de palabras/glosas	3071	2876
Validación	Pares de Frases	50	
	Nº de palabras/glosas	582	574
Test	Pares de Frases	50	
	Nº de palabras/glosas	570	505

C. Resultados de los experimentos realizados

En primer lugar se muestran los resultados de tres experimentos realizados para el sistema de reconocimiento de voz: En el *Experimento 1* se evalúa el sistema de reconocimiento con el modelo de lenguaje y el vocabulario generados a partir de un conjunto de entrenamiento. Esta situación es la más realista, y sus resultados fueron: WER=*24,08, I=2,61%, B=6,71%, S=14,76%*. En el *Experimento 2* el modelo de lenguaje se genera a partir del conjunto de entrenamiento, mientras que el vocabulario incluye todas las palabras (entrenamiento y test). Sus resultados fueron: *WER=15,84, I=1,19%, B=5,93%, S=8,72%*. Finalmente, en el *Experimento 3* se utilizan todas las frases para el entrenamiento y para el vocabulario. Se intenta conseguir que se disponga de tantas frases de entrenamiento que las frases de test estén contenidas en ellas. Sus resultados fueron: *WER=4,74, I=0,86%, B=1,94%, S=1,94%.*

En la tabla II se observan los resultados de los experimentos de traducción realizados, tanto con las frases de referencia del corpus paralelo castellano-LSE ("Ref"), como utilizando la salida del reconocedor de voz para los tres experimentos de reconocimiento comentados anteriormente ("Exp" 1-3). Por otro lado, se muestran los resultados

habiendo entrenado el modelo de traducción con las frases de referencia, y los mismos resultados considerando la salida de reconocedor (de las frases de entrenamiento) para entrenar el modelo de traducción.

TABLA II
RESULTADOS DE LOS EXPERIMENTOS DE TRADUCCIÓN

Modelo de traducción generado con las frases de referencia del conjunto de entrenamiento				
		WER	BLEU	NIST
Traducción basada en subfrases	Exp 1	39,17	0,4853	6,2806
	Exp 2	37,99	0,4900	6,4006
	Exp 3	33,72	0,5301	6,7238
	REF	31,75	0,5469	6,8652
Traducción basada en FST	Exp 1	35,85	0,5090	6,6473
	Exp 2	33,89	0,5238	6,8610
	Exp 3	**29,32**	**0,5804**	**7,3100**
	REF	**28,21**	**0,5905**	**7,3501**
Modelo de traducción generado con la salida del reconocedor para las frases de entrenamiento				
		WER	BLEU	NIST
Traducción basada en subfrases	Exp 1	40,04	0,4775	6,2076
	Exp 2	37,46	0,4939	6,4738
	Exp 3	32,44	0,5449	6,8606
	REF	31,75	0,5469	6,8652
Traducción basada en FST	Exp 1	36,33	0,5188	6,5273
	Exp 2	33,42	0,5235	6,8344
	Exp 3	**29,27**	**0,5698**	**7,1953**
	REF	**28,21**	**0,5905**	**7,3501**

Los resultados de la Referencia siempre serán los mejores resultados (menor WER y mayor BLEU y NIST) en comparación con los obtenidos en la traducción de la salida del reconocedor de voz puesto que la referencia no contiene errores de reconocimiento que dificultan la traducción. Además, cuanto peor es la tasa de reconocimiento, peor es la tasa de traducción que se consigue traduciendo la salida del reconocedor. En general, con esta base de datos, la traducción estadística basada en FST ofrece mejores resultados entrenando con las salidas del reconocedor. Se observa también que al entrenar el modelo de traducción con las salidas de reconocimiento se entrena dicho modelo con los posibles errores del reconocedor, de forma que el modelo de traducción puede aprender de estos errores y corregirlos durante el proceso de traducción.

VI. CONCLUSIONES

En este artículo se ha presentado un sistema de traducción estadística de voz en castellano a lengua de signos para personas sordas. Las soluciones tecnológicas estudiadas utilizan un modelo de traducción basado en subsecuencias de palabras y un transductor de estados finitos (FST). Los resultados mostrados corresponden a pruebas con el texto original (de referencia) y el texto obtenido a la salida del reconocedor. Entrenar el modelo de traducción con las salidas de reconocimiento permite aprender de los errores y corregirlos durante la traducción. Finalmente se puede concluir que el mejor sistema es el de la traducción basada en FST entrenando con las salidas del reconocedor con una WER de 29,27% y un BLEU de 0,5698, para la base de datos utilizada, con respecto al modelo que utiliza subsecuencias de palabras para la traducción.

VII. PROTOTIPO DESARROLLADO

Con este trabajo se ha desarrollado un prototipo (Fig. 4) de traducción de voz a LSE que ha sido evaluado con frases pronunciadas por estudiantes. El siguiente paso es evaluar el sistema en condiciones reales considerando interacciones reales entre los policías y personas sordas.

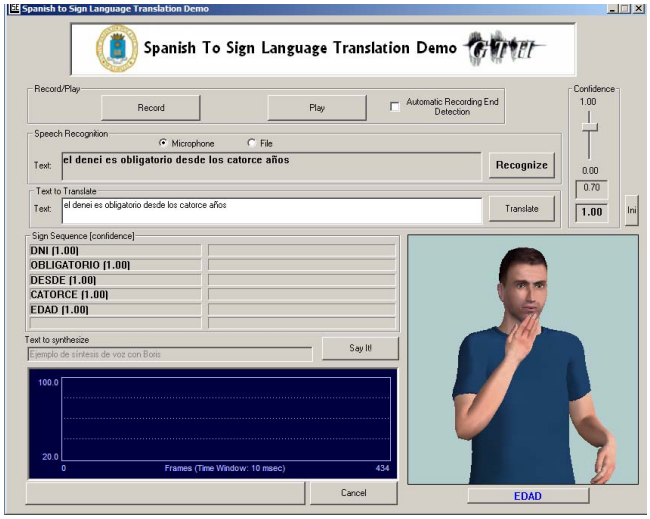

Figura 4: Interfaz del prototipo.

REFERENCIAS

[1] http://www.sign-lang.unihamburg.de/eSIGN
[2] Stolcke A. "SRILM – An Extensible Language Modelling Toolkit" Intl. Conf. on Spoken Language Processing, 2002.
[3] Och J., Ney H. "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, Octubre 2000.
[4] http://www.statmt.org/moses
[5] Vidal E., Casacuberta F, García P. "Gramatical Inferrence and Automatic Speech Recognition". *New Advances and Trends in Speech Recognition and Coding* (volume 147 of NATO-ASI Series F: Casacuberta and Vidal
[6] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311-318.
[7] http://www.nist.gov/speech/tests/mt/

Beatriz Gallo (bgallo@die.upm.es). Beatriz realizó su trabajo fin de carrera (2007) en la Universidad Politécnica de Madrid obteniendo la máxima calificación. Su Proyecto fin de carrera se ha centrado en la traducción de voz a Lengua de Signos con herramientas estadísticas.

Rubén San-Segundo (lapiz@die.upm.es). Rubén realizó su trabajo fin de carrera (1997) y su tesis doctoral (2002) en la Universidad Politécnica de Madrid obteniendo la máxima calificación en ambos trabajos. Durante los años 1999 y 2000, Rubén realizó dos estancias en The Center of Spoken Language Research (CSLR), University of Colorado (Boulder). Desde septiembre de 2001 hasta febrero de 2003 Rubén trabajó en la División de Tecnología del Habla de Telefónica I+D.
Rubén es autor y coautor de gran cantidad de artículos internacionales en tecnología del habla. Es coordinador de la Red Temática en Tecnologías del Habla y miembro del SIG-IL (Grupo de Interés de la ISCA en Iberian Languages) Board. Actualmente, Rubén es profesor del Departamento de Ingeniería Electrónica de la E.T.S.I. de Telecomunicación de la Universidad Politécnica de Madrid y miembro del Grupo de Tecnología del Habla.

Juan Manuel Lucas-Cuesta (juanmak@die.upm.es). Juan Manuel es Ingeniero de Telecomunicaciones por la Universidad Politécnica de Madrid (26-10-2006). Recibió el Diploma de Estudios Avanzados en la Universidad Politécnica de Madrid (10-07-08). En la actualidad está realizando la Tesis Doctoral en el Grupo de Tecnología del Habla de dicha Universidad, bajo la dirección de Javier Ferreiros. De julio a septiembre de 2009 realizó una estancia breve de investigación en la University of Ulster, en el campus de Jordanstown, Irlanda del Norte (UK).

Roberto Barra Chicote (barra@die.upm.es). Roberto recibió el título de Ingeniero de Telecomunicación por la Universidad Politécnica de Madrid en 2005. Desde 2003 es miembro del Grupo de Tecnología del Habla. En 2006 fue investigador visitante en el Center for Spoken Language Research (CSLR) en la Universidad de Colorado. En 2008 fue investigador visitante en Centre for Speech Technology Research (CSTR) en la Universidad de Edimburgo. Su labor investigadora se centra en síntesis de habla con contenido emocional y reconocimiento automático de emociones.

Luis Fernando D'Haro (lfdharo@die.upm.es). Luis Fernando D'Haro recibió el título de Doctor por parte de la Universidad Politécnica de Madrid (UPM) en el año 2009, y el título de Ingeniero Electrónico por parte de la Universidad Autónoma de Occidente en Cali, Colombia en el año 2000. Obteniendo en ambas ocasiones las máximas distinciones. Es profesor asistente en la UPM (Madrid, España) desde el año 2007. En el año 2005 realizó una estancia de investigación en el Computer Science VI, RWTH de la Universidad de Aachen (Alemania) trabajando en el área de traducción automática y modelado de lenguaje. En 2006 realizó otra estancia en los laboratorios de investigación de AT&T en Florham Park, NJ (EEUU), trabajando en interacción multimodal e interfaces. Su investigación se enfoca en sistemas de diálogo multimodales y el modelado de lenguaje.

Fernando Fernández Martínez (ffm@die.upm.es). Fernando recibió los títulos de Ingeniero de Telecomunicación (2002) y Doctor Ingeniero de Telecomunicación (2009) por la Universidad Politécnica de Madrid (UPM), con la mayor de las distinciones ("Cum Laude"). Durante 2006, Fernando realizó una estancia de investigación en el IDIAP Research Institute afiliado a la â€œEcole Polytechnique Fédérale de Lausanneé (EPFL) y la Universidad de Ginebra (Suiza). Desde Diciembre del 2004 ocupa una plaza como Profesor Ayudante en el Departamento de Ingeniería Electrónica de la UPM. La actividad investigadora del Dr. Fernández gira en torno a la Tecnología del Habla (especialmente en los Sistemas de interacción o diálogo Persona-Máquina).