

# Sistema de Traducción Estadística de Voz a Lengua de Signos para Personas Sordas

B. Gallo, R. San-Segundo, J.M. Lucas, R. Barra, L.F. D'Haro, F. Fernández  
Grupo de Tecnología del Habla. Universidad Politécnica de Madrid.  
lapiz@die.upm.es

Ponencia

**Abstract** — Este artículo presenta un conjunto de experimentos para la realización de un sistema de traducción estadística de voz a lengua de signos para personas sordas. Éste contiene un módulo de reconocimiento de voz, uno de traducción estadística de palabras en castellano a signos en Lengua de Signos Española, y un tercer módulo que representa los signos mediante un avatar. La traducción se hace mediante dos alternativas tecnológicas: la primera basada en modelos de subsecuencias de palabras y la segunda basada en transductores de estados finitos. Para la evaluación se utilizan varias métricas, como WER (tasa de error de palabras), BLEU y NIST. Estas pruebas incluyen experimentos con las frases originales en castellano y Lengua de Signos y con frases extraídas del reconocedor de voz. Se evalúan también diferentes situaciones del reconocedor de voz a la hora de obtener el vocabulario y el modelo de lenguaje. Los mejores resultados se obtienen con la traducción mediante transductores de estados finitos dando unas tasas de error de 28,21% para las frases de referencia, y de 29,27% la salida del reconocedor.

## 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.

## 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:

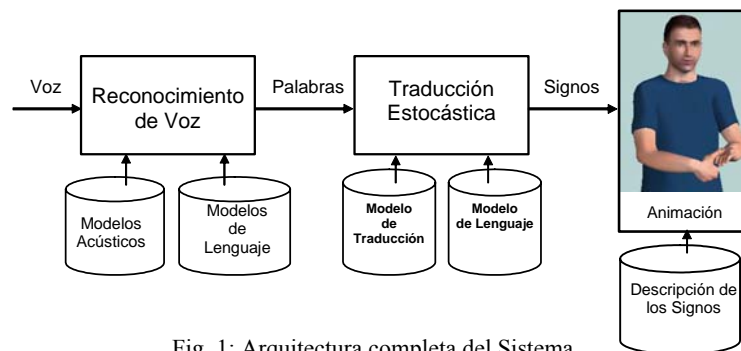


Fig. 1: Arquitectura completa del Sistema

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].

Las arquitecturas de las dos soluciones propuestas para la traducción estadística son:

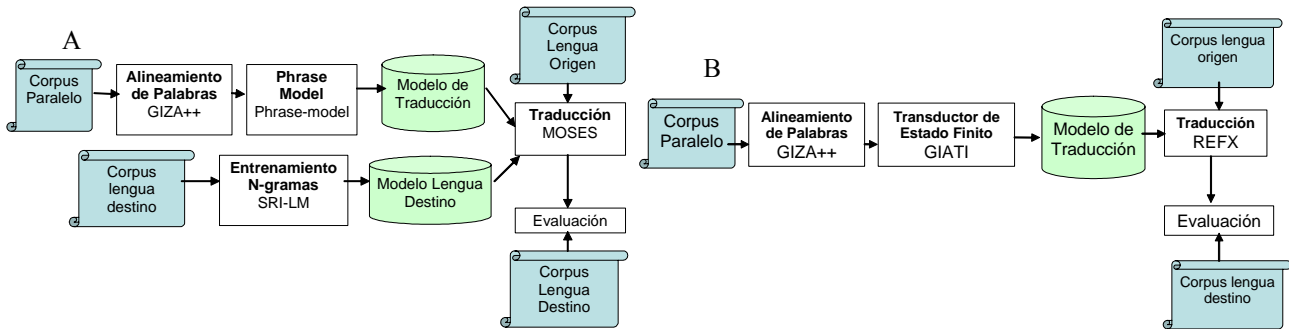


Fig. 2: (A) Arquitectura completa de la Traducción basada en Subsecuencias de palabras. (B) Arquitectura completa de la Traducción basada en Transductores de Estados Finitos. Todas las herramientas utilizadas para la traducción son de software libre.

### 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 figura 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  $p(o|d) \cdot p(d)$ , 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 figura 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:

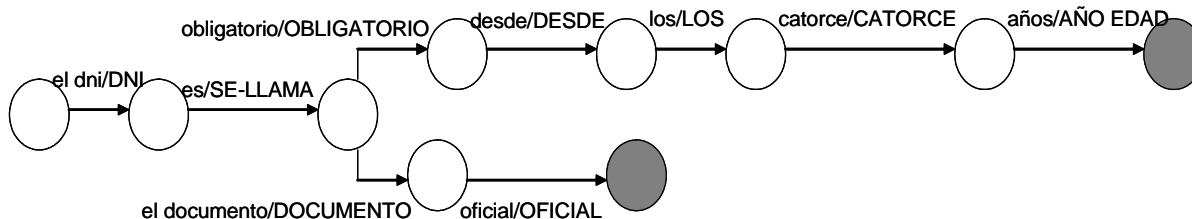


Fig. 3: Transductor de estado finito inferido a partir del bigrama del ejemplo anterior

#### VI. 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 que diría un funcionario en una comisaría. El conjunto de frases se dividió aleatoriamente en tres grupos: entrenamiento (conteniendo el 70% de las frases), evaluación (15%) y test (15%).

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 I 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 I  
RESULTADOS DE LOS EXPERIMENTOS DE TRADUCCIÓN

		Modelo de traducción generado con las frases de referencia del conjunto de entrenamiento						Modelo de traducción generado con la salida del reconocedor para las frases de entrenamiento					
		WER	I(%)	B(%)	S(%)	BLEU	NIST	WER	I(%)	B(%)	S(%)	BLEU	NIST
Traducción basado en subsecuencias	Exp 1	39,17	13,33	12,04	13,8	0,4853	6,2806	40,04	14,42	12,44	13,18	0,4775	6,2076
	Exp 2	37,99	12,35	11,61	14,03	0,49	6,4006	37,46	12,25	11,76	13,45	0,4939	6,4738
	Exp 3	33,72	12,39	9,82	11,51	0,5301	6,7238	32,44	11,63	9,28	11,53	0,5449	6,8606
	Ref	31,75	11,98	8,93	10,84	0,5469	6,8652	31,75	11,98	8,93	10,84	0,5469	6,8652
Traducción basada en FST	Exp 1	35,85	9,15	14,34	12,36	0,509	6,6473	36,33	11,48	12,61	12,24	0,5188	6,5273
	Exp 2	33,89	8,62	13,47	11,8	0,5238	6,861	33,42	9,00	13,18	11,24	0,5235	6,8344
	Exp 3	<b>29,32</b>	<b>9,00</b>	<b>11,92</b>	<b>8,40</b>	<b>0,5804</b>	<b>7,31</b>	<b>29,27</b>	<b>7,65</b>	<b>11,17</b>	<b>10,45</b>	<b>0,5698</b>	<b>7,1953</b>
	Ref	<b>28,21</b>	<b>8,41</b>	<b>11,05</b>	<b>8,75</b>	<b>0,5905</b>	<b>7,3501</b>	<b>28,21</b>	<b>8,41</b>	<b>11,05</b>	<b>8,75</b>	<b>0,5905</b>	<b>7,3501</b>

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.

## VII. 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.

## AGRADECIMIENTOS

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).

## REFERENCIAS

- [1] <http://www.sign-lang.uni-hamburg.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 Inference 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/>