HAL
open science

Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web

Luis Villaseñor-Pineda, Manuel Montes-Y-Gómez, Dominique Vaufreydaz, Jean-François Serignat

To cite this version:

Luis Villaseñor-Pineda, Manuel Montes-Y-Gómez, Dominique Vaufreydaz, Jean-François Serignat. Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web. Internationnal Conference on Computing (CIC-2003), Oct 2003, Mexico City, México. 3 p. inria-00326513

HAL Id: inria-00326513

https://inria.hal.science/inria-00326513v1

Submitted on 3 Oct 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web

L. Villaseñor-Pineda‡, M. Montes-y-Gómez‡, D. Vaufreydaz* & J-F. Serignat*

‡ Laboratorio de Tecnologías del Lenguaje, Ciencias Computacionales, INAOE, México
{villasen,mmontesg}@inaoep.mx

*Laboratoire CLIPS/IMAG, Francia
{Dominique.Vaufreydaz,Jean-Francois.Serignat}@imag.fr

RESUMEN

En este artículo presentamos una metodología para la elaboración de un corpus balanceado fonéticamente para el español mexicano. Este corpus será utilizado para el entrenamiento y evaluación de modelos acústicos indispensables en el proceso de reconocimiento del habla. En la primera parte de este artículo se explica la motivación de este trabajo. Posteriormente, se explica el método utilizado y una serie de consideraciones en particular para el lenguaje español. Por último, se describen las características del corpus obtenido y se compara su distribución fonética con otros estudios del lenguaje español tanto latinoamericano como ibérico.

Palabras clave:
Recolección de corpus, distribución fonética, reconocimiento del habla.

I. INTRODUCCIÓN

El presente trabajo se inscribe dentro el proyecto "Interacción Oral Hombre-Máquina", uno de cuyos objetivos es la construcción de un reconocedor de habla para el español mexicano. Entre los elementos necesarios para la construcción de este reconocedor es necesario contar con una colección de grabaciones, la cual servirá de base para el cálculo de los modelos acústicos pertinentes. Dicha colección de grabaciones deberá cuidar ciertos aspectos para que el reconocedor sea lo más robusto posible. Dos de estos aspectos son abordados por el presente trabajo: (i) el corpus oral debe ser rico, es decir debe contener todos los fonemas del español mexicano, y (ii) debe ser balanceado, es decir, debe conservar la distribución fonética del español mexicano.

El primer paso para la elaboración de un corpus rico y balanceado es la selección de un conjunto de frases, las cuales serán posteriormente grabadas bajo condiciones controladas. El enfoque tradicional involucra un enorme esfuerzo en la selección de las frases pertinentes. Primero, es necesario realizar una selección minuciosa de palabras ricas fonéticamente, para después, elaborar frases a partir de dichas palabras. Posteriormente, es necesario verificar la distribución fonética de las frases resultantes, en caso de despegarse en demasía de la realidad es necesario eliminar unas cuantas frases para agregar otras que mejoren la distribución. Por supuesto, el cambio de unas frases por otras afectará la distribución de otros fonemas y, repetiremos cuantas veces sea necesario, el paso anterior hasta alcanzar una distribución apropiada.

El método propuesto en este trabajo difiere substancialmente del enfoque tradicional. Éste retoma y adapta la metodología propuesta por [1] para el idioma francés. La idea principal de este método para la elaboración de un corpus fonéticamente balanceado recae en la hipótesis de que la Web, por su enorme tamaño, ya es una fuente rica y balanceada fonéticamente. Por lo tanto, bastará con recuperar un conjunto de frases cualesquiera (a las que sólo será necesario aplicar una serie de filtros para facilitar su lectura). Este método se describe en la siguiente sección; y posteriormente, se compara el corpus obtenido con otros estudios sobre el español, confirmando la hipótesis de partida.

II. RECOLECTANDO DOCUMENTOS DE LA WEB

El primer paso para la elaboración del corpus consistió en el acopio de documentos de Internet (para una exposición completa de esta problemática véase [2, 3]). Para ello fue utilizado un robot encargado de recolectar páginas Web. Este robot, desarrollado en el laboratorio CLIPS [1], toma una o varias direcciones y a partir de ellas recoge todas las páginas y documentos de texto. El robot filtra los documentos de acuerdo a su tipo, separando documentos html y textuales de imágenes, audio, etc. y, de acuerdo al nombre del servidor Web, limitando los dominios de interés.

Dado que los textos recolectados de la Web no están en un formato adecuado para ser leídos por un locutor y así ser grabados, es necesario aplicar una serie de filtros para obtener el formato apropiado. El primer paso es la eliminación de etiquetas, encabezados y demás información incrustada en los documentos necesaria para su correcto manejo por los navegadores de Internet. Después de aplicar este primer filtro el tamaño de la colección resultante fue de 1.2 Gbytes, con un total de 244,251,605 de palabras y un total de líneas de 15,081,123.

El segundo paso es asegurarnos que contamos con documentos en español. A pesar de restringir la colecta de documentos a dominios hispanohablantes, es normal encontrar documentos en otros idiomas. Dado que estamos interesados en recuperar frases y no documentos enteros, se filtraron los documentos usando un léxico exclusivamente del español. Esto implico la rea-

lización de un trabajo previo, dados los limitados recursos lingüísticos disponibles para el español mexicano, en particular la carencia de un léxico lo suficientemente rico. A continuación se describe brevemente el trabajo realizado para la recopilación de un léxico para el español.

III. CONSTRUYENDO UN LÉXICO PARA EL ESPAÑOL

El léxico requerido para este segundo filtro consiste en todas las palabras posibles del español incluyendo todas sus inflexiones. Es decir, necesitamos un léxico con el mayor número posible de formas léxicas. El método propuesto consistió en utilizar un léxico inicial tomado de un conjunto de documentos que presumiblemente contenían sólo palabras en español (un conjunto de artículos tomados de periódicos y revistas mexicanas, y de dos diccionarios del idioma español). Con ello logramos un léxico inicial de poco más de 177,000 vocablos. Con este léxico se inicio el proceso de filtrado y se obtuvo un conjunto de frases, llamado Corpus170. En la sección de resultados se encuentra la comparación del Corpus170 con otros estudios describiendo su distribución fonética.

A pesar de que los resultados obtenidos fueron excelentes se realizó otro experimento para observar la importancia del léxico sobre el corpus recolectado, ya que el léxico usado fue relativamente pequeño. Para ello se modificó el proceso de recolección de frases. Este proceso consistió en filtrar dos veces el mismo grupo de documentos. En la primera ocasión, al usar el léxico original, se obtuvieron las palabras desconocidas. Para diferenciar las palabras en español dentro de esta lista de vocablos desconocidos nos apoyamos nuevamente en la Web. Gracias a las herramientas proporcionadas por el buscador Google[1] se programó una herramienta que obtenía la frecuencia de la palabra en la Web en documentos en español. Si la frecuencia era mayor de cierto umbral la palabra era añadida al léxico. Con este nuevo léxico enriquecido se filtró por segunda vez la misma colección de documentos consiguiendo un conjunto mayor de frases. Este segundo corpus lo llamamos Corpus230.

Por último, un tercer filtro es aplicado para eliminar las frases que cumplan alguno de los siguientes criterios:

- Frases de menos de 30 palabras
- Frases duplicadas
- Frases incluyendo más de un "."
- Frases incluyendo dos palabras consecutivas idénticas

La Tabla 1 compara los datos descriptivos del Corpus170 y del Corpus230. En la siguiente sección se discute sobre la pertinencia del corpus elaborado al verificar su distribución fonética.

[1] http://www.google.com.mx/intl/es/options.html

	Tamaño del léxico usado	Número de frases	Número de palabras	Promedio de palabras por frase
Corpus170	177,290	339,833	14,511,061	42.7
Corpus230	235,891	344,619	14,766,638	42.8

Tabla 1. Datos descriptivos de las colecciones Corpus170 y Corpus230.

IV. EVALUACIÓN DEL CORPUS

Dado que las grabaciones obtenidas serán dedicadas al entrenamiento de modelos acústicos para el reconocimiento del habla, es necesario verificar si la distribución de la frecuencia de los fonemas es correcta en nuestro corpus. Para ello, se efectúo una comparación de nuestra distribución con la distribución del español reportada en la literatura. La Figura 1 muestra el comparativo de nuestra distribución (Corpus170 y Corpus230) con un estudio del español latinoamericano y con un estudio del español ibérico[2]. Por supuesto, fue necesaria la transformación de cada palabra del corpus en sus fonemas correspondientes, para ello usamos la herramienta desarrollada en el proyecto DIME [4].

Como primera observación se puede confirmar la enorme coincidencia entre el Corpus170 y el Corpus230. A pesar de la diferencia en tamaño de ambas colecciones, éstas presentan casi la misma distribución con un coeficiente de correlación de 0.99. En cierta manera, este resultado no es sorprendente, pues a pesar de incrementar el léxico en casi 60 mil formas léxicas más (un incremento del 33.1%), el número de frases incrementado fue de poco menos de 5,000 frases (un incremento del 1.5%).

Por otro lado, como conclusión principal podemos afirmar la hipótesis inicial, ya que la distribución de fonemas observada en las frases recolectadas de la Web es muy cercana a las distribuciones reportadas para el español. Los coeficientes de correlación entre el Corpus230 y los estudios latinoamericano [5] e ibérico [6] son de 0.994 y de 0.942 respectivamente. Como observación interesante, cabe notar el caso de los fonemas /a/ y /e/. En el estudio del español ibérico la frecuencia de la /a/ es mayor a la del fonema /e/. En el caso del español latinoamericano esta relación es inversa. En nuestro corpus, estas frecuencias son más cercanas al estudio latinoamericano.

V. TRABAJO FUTURO

Antes de iniciar el proceso de grabación será necesario reducir el tamaño del Corpus230. De hecho sólo serán

[2] Debido a que los fonemas son unidades abstractas existen diversas formas de representarlos, para efectos de esta exposición, hemos optado por usar una notación utilizando algunos de los símbolos que en el alfabeto corriente más o menos corresponden con dichas entidades.

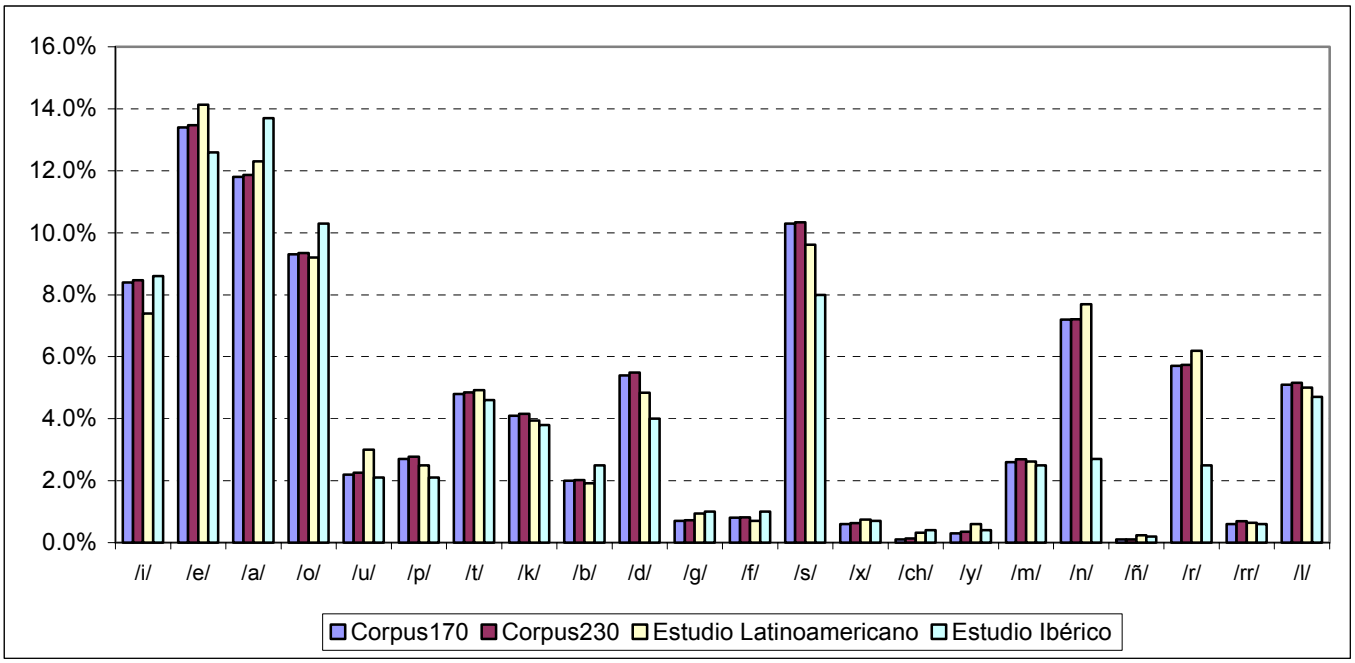

Figura 1. Comparación de la distribución porcentual de la frecuencia de los fonemas en el español.

necesarias del orden de 6000 frases suficientes para construir una base de datos con 10 horas de grabación. Una forma automática para lograr esta reducción se basa en el cálculo de la perplejidad de la oración a través de un modelo de lenguaje. Así, después de calcular la perplejidad para cada frase, podríamos conservar las seis mil frases con menor perplejidad. Actualmente se trabaja en la construcción de un modelo de lenguaje para este propósito. Dado que nuestra intención final es usar el reconocedor de voz resultante en situaciones de habla espontánea hemos iniciado la construcción de este modelo usando una colección de transcripciones estenográficas de conversaciones entre varios individuos.

Por último, una última fase será la revisión manual de cada frase para juzgar sobre su contenido. Frases con contenido sexual o hilarante deben excluirse para evitar posibles complicaciones durante la grabación que impactarían sobre la calidad y/o el tiempo de efectivo de grabación.

AGRADECIMIENTOS

El desarrollo de este trabajo fue parcialmente financiado por el Laboratorio Franco-Mexicano de Informática (LAFMI) bajo el marco del proyecto “Interacción oral hombre-máquina”. Los autores también desean agradecer a la M. C. Esmeralda Uraga por permitir el uso de su sistema de transformación fonética. Así como al programa de *Verano de la Investigación Científica* de la Academia Mexicana de Ciencias, que permitió la estancia de los estudiantes Andrés González y Alberto López en el Laboratorio de Tecnologías del Lenguaje. Ambos estudiantes participaron activamente en la programación de las herramientas usadas en el presente trabajo.

REFERENCIAS

[1] D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, M. Akbar, *A New Methodology For Speech Corpora Definition From Internet Documents*, LREC'2000 (Language Resources & Evaluation international Conference), Athens (Greece), pp. 423-426, June 2000.

[2] S. Galicia-Haro. *Procesamiento de Textos Electrónicos para la Construcción de un Corpus*. CORE-2003, México, D.F. 2003.

[3] Gelbukh, A., G. Sidorov and L. Chanona. Compilation of a Spanish Representa-tive Corpus. *International Conference on Computational Linguistics and Intelligent Text Processing CICLing02,* LNCS 2276, Springer. 2002.

[4] E. Uraga and L. Pineda. Automatic generation of pronunciations lexicons for Spanish, *Computational Linguistics and Intelligent Text Processing. CICLing 2002*. LNCS 2276, Springer, 2002.

[5] H. E. Pérez. Frecuencia de fonemas. *Revista Electrónica de la Red Temática en Tecnologías del Habla*, Número 1, Marzo, 2003, ISSN: 1695-9914

[6] E. Alarcos-Llorach. *Fonología española*. Madrid, Gredos. 1965.