6

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

UN SISTEMA DE RECONOCIMIENTO DE VOZ PARA EL ESPAÑOL

TESIS PROFESIONAL QUE PARA OBTENER

EL TÍTULO DE:

LICENCIADO EN INFORMÁTICA

P R E S E N T A:

CÉSAR FRANCISCO GAMBOA VERDUZCO

ASESORA:

M.C. ESMERALDA URAGA SERRATOS

MÉXICO, D.F. 2002

Agradecimientos:

A la M. C. Esmeralda Uraga, por su amistad, paciencia y dedicación para realizar este trabajo de tesis.

A Carolina Sazuelta y Haydé Castellanos de la Facultad de Filosofía y Letras de la UNAM, muchas gracias por su apoyo en la redacción y edición de esta tesis.

Al grupo de Inteligencia artificial del IIMAS, Luis Pineda, Roxana Philips, Javier Cuétara, Paulino Ochoa, Arturo Espinosa e Iván Meza, gracias por su constante apoyo.

A CONCAYT (convenio 27948-A) por el apoyo para desarrollar este trabajo de tesis.

Dedicatorias:

A mi familia y amigos.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la UNAM se desarrolla el proyecto *Diálogos Inteligentes Multimodales en Español* (DIME), cuyo objetivo es la creación de un sistema computacional que interactúe por medio del lenguaje hablado y de una interfaz gráfica con el usuario [Pineda, 2001].

El trabajo del proyecto DIME se enfoca actualmente en tres tareas principales:

1. *Grabación y transcripción de diálogos.* El *corpus* del proyecto DIME (corpus DIME) está formado por un conjunto de grabaciones de audio y video con sus correspondientes transcripciones ortográficas. Este corpus registra la interacción y conversación de dos personas realizando el diseño[1] de una cocina (ver figura 1.1). Los fenómenos lingüísticos observados en los diálogos de este *corpus* son muy importantes para modelar computacionalmente como hablan e interactúan las personas.

2. *Construcción de un sistema de reconocimiento de voz.* Este sistema debe reconocer las palabras que pronuncie cualquier persona que hable Español (ver figura 1.2), aunque el vocabulario de reconocimiento sólo contiene palabras relacionadas al dominio de diseño de cocinas.

3. *Crear una gramática del Español.* Esta gramática debe describir como se forman las frases en Español. También debe contemplar los fenómenos sintácticos de las frases observadas en el *corpus*. A partir del análisis gramatical se obtiene el significado de las frases y su interpretación en términos de la aplicación del sistema. ésto permite que el sistema actúe y responda de manera adecuada.

[1] El diseño se realizó con el programa *Home Designer* creado por *ALPHA Software.* http://www.alpha.software.com

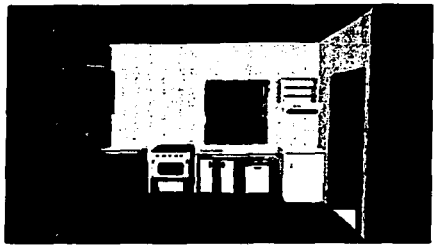


Figura 1.1: Diseño gráfico de una cocina.

Figura 1.2: Interacción humano - máquina por medio del habla.

De manera específica, este trabajo de tesis contribuye al desarrollo del módulo de reconocimiento de voz (ver figura 1.3). Este módulo utiliza tres modelos principales: un modelo acústico, un modelo de lenguaje y un modelo de pronunciación. El objetivo general de este trabajo de tesis es crear un nuevo modelo acústico para el Español que mejore el desempeño del sistema actual. Un modelo acústico permite identificar la secuencia de sonidos que emite una persona al hablar.

Un modelo acústico se construye a partir de un conjunto de grabaciones de voz. Las grabaciones se procesan con el objetivo de extraer las características acústicas de los sonidos de voz. Estas características son los datos de entrada utilizados por métodos estadísticos, a partir de los cuales, se crean los modelos acústicos. Los modelos acústicos creados en este trabajo de tesis están basados en modelos ocultos de Markov y fueron implementados utilizando la herramienta *Hidden Markov Model Tool Kit* (**HTK**).

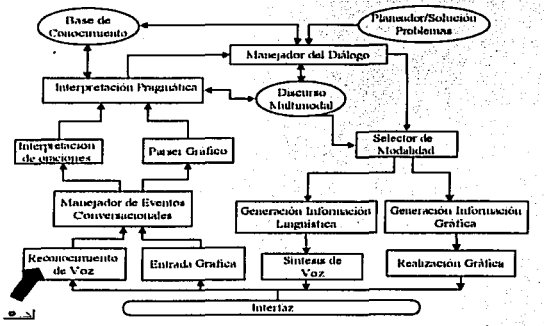


Figura 1.3: Arquitectura del sistema que se desarrolla en el proyecto DIME. En la parte inferior izquierda se encuentra señalado el módulo correspondiente al reconocimiento de voz.

1.1 Antecedentes

En México, algunos grupos trabajan desarrollando sistemas de reconocimiento de voz para el español hablado en México. Uno de estos grupos es Tlatoa, de la UDLA Campus Puebla, el cual ha realizado algunos prototipos para reconocimiento de voz independiente del hablante [Tlatoa, 1998]. Algunos prototipos que ha desarrollado Tlatoa son: un sistema de correo electrónico y de voz usando reconocimiento de voz [Munive,1999], un reconocedor fonético de dígitos, [Munive, 1999], una aplicación para la enseñanza del Español [Kirschning, 2000], entre otros.

Existe otro grupo en el ITESM Campus Morelos que ha trabajado en el desarrollo de aplicaciones y modelos para reconocimiento de voz. Algunos prototipos que este grupo desarrolló son: un reconocedor de dígitos basado en redes neuronales [Garcia, 1994], una aplicación para solicitar la transferencia de llamadas telefónicas por medio de voz [Uraga, 1999], entre otros. La mayoría de las aplicaciones que han desarrollado estos grupos se realizaron con la herramienta *CSLU Toolkit* [CSLU, 2002]. Estas aplicaciones utilizan un vocabulario restringido y están limitadas a un dominio muy específico.

Actualmente, en el grupo de inteligencia artificial del IIMAS se desarrolla un módulo de reconocimiento de voz para el sistema del proyecto DIME [Pineda, 2001]. Este módulo permite establecer un diálogo hablado en el diseño de cocinas. El

desempeño actual de este módulo es del 50% de reconocimiento correcto de palabras utilizando voz continua. En este trabajo de tesis se crearán nuevos modelos acústicos más robustos para mejorar este módulo de reconocimiento de voz.

En la siguiente sección se describen las razones por las que la tarea de reconocimiento de voz resulta difícil.

1.2 Planteamiento del problema

El problema general que se aborda en este trabajo de tesis es el de reconocimiento de voz continua e independiente del hablante para el Español hablado en México. La voz continua es la forma natural de hablar sin hacer pausas obligatorias entre la pronunciación de una palabra y otra [Uraga, 1999]. La independencia del hablante se refiere a que, independientemente del tipo de voz de la persona que hable, las palabras pronunciadas deben ser identificadas. Esta propiedad permite que cualquier persona interactúe con el sistema utilizando el habla como medio de comunicación.

Estas son las razones por las cuales la tarea de reconocimiento de voz es difícil [Yannakoudakis, 1985]:

- La voz continua tiene que ser segmentada para obtener la información acústica correspondiente a cada sonido emitido en el habla.
- Elementos individuales como fonemas o palabras tienden a perder su propia identidad en el habla; es decir, algunos elementos sufren efectos de coarticulación. Por ejemplo: en la frase "*es tan bella*", la pronunciación de /*n*/ se realiza como /*m*/.
- Las muestras de voz varían no solo entre hablantes, también en la pronunciación de una misma palabra pronunciada diferentes veces por un mismo hablante.
- Una palabra puede variar en tono, acentuación y en la velocidad de pronunciación.
- Dependiendo del origen geográfico de los hablantes las palabras pueden ser pronunciadas de distinta forma.
- Las pronunciaciones de diferentes palabras pueden parecerse demasiado. Por ejemplo: *desarrollo* y *desarrolló* varían en la acentuación, *inferior* e *interior* varían en un sólo sonido.
- Ruido ambiental y otro tipo de interferencias pueden distorsionar la señal original y hacer que los sonidos del habla no se entiendan.

El problema específico que se trató en este trabajo se plantea a continuación:

A partir del sonido generado al pronunciar una secuencia de palabras, el problema consiste en identificar las unidades fonéticas correspondientes (ver figura 1.4).

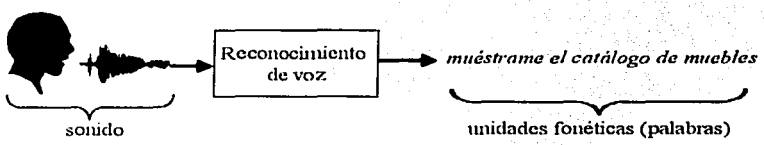


Figura 1.4: Ilustración del proceso de reconocimiento de voz.

1.3 Objetivos

1.3.1 Objetivo general

El objetivo general de este trabajo de tesis es modelar acústicamente la voz del español hablado en México, utilizando un enfoque basado en modelos ocultos de Markov, con el propósito de construir un sistema de reconocimiento de voz continua e independiente del hablante.

1.3.2 Objetivos específicos

Los objetivos específicos de este trabajo de tesis son los siguientes:

- Plantear y seguir una metodología para obtener un *corpus* de voz con el objetivo de crear modelos acústicos de las unidades fonéticas del Español.
- Crear un conjunto de modelos acústicos - fonéticos, a partir del *corpus*, utilizando programas de la herramienta HTK [Young, 1997]. Cada modelo tendrá las siguientes características:
  - La unidad que modela es el fonema.
  - Un entrenamiento basado en modelos ocultos de Markov.
  - Un modelado independiente del contexto.
- Integrar los modelos acústicos a un sistema de reconocimiento de voz.

- Evaluar el desempeño de los modelos acústicos - fonéticos en el sistema de reconocimiento de voz continua.

1.4 Importancia del trabajo

Por medio de los sistemas de reconocimiento de voz se tiene una nueva forma de interactuar con las máquinas.

El uso de tecnología con reconocimiento de voz puede brindar beneficios en diferentes aspectos (ejemplo: social, comercial, etc.). El campo de las telecomunicaciones, la telefonía digital, la inteligenica artificial y la robótica han tenido un gran avance tecnológico y con ello se están convirtiendo en medios cada vez más accesibles y populares, a su vez ha surgido una nueva forma más natural de controlar algunos de estos dispositivos y es por medio de la voz, que al ser usada para controlar máquinas trae otros beneficios como mejorar la productividad de los procesos industriales, la eficacia de los servicios médicos, facilitar los trabajos de exploración submarina o minera, la manipulación de sustancias químicas, etc.; tareas de la que pueden ser liberados los humanos, por incrementar su eficacia o para reducir riesgos de trabajo.

Aunque muchos de estos sistemas todavía están en desarrollo, tienen el potencial de revolucionar la forma en que la gente interactúa con las computadoras. En un futuro, los sistemas de lenguaje hablado serán capaces de permitir la interacción humano - máquina en una forma natural que no requerirá de entrenamiento especial. Estas interfaces permitirán que los recursos basados en computadora sean disponibles a grupos nuevos de usuarios (usuarios casuales, usuarios discapacitados, usuarios con un lenguaje diferente al nativo). También se permitirá apoyar a usuarios expertos en el manejo de problemas de información intensiva y proporcionar diferentes servicios a varias clases de usuarios (usuarios con ojos o manos ocupadas, usuarios novatos o discapacitados) proporcionando una modalidad conveniente y natural de accesar y manipular información y sistemas en general [Cole, 1995].

Con el objetivo de incluir a todo tipo de usuarios, actualmente se desarrollan páginas web que permiten al usuario interactuar con el sistema mediante la voz. Esto implica que cada web cuente con una versión preparada para que resulte más sencillo navegar a personas con algún tipo de discapacidad física.

En el ámbito comercial, una empresa que se beneficia con este tipo de tecnología es *Royal Phillips Electronics*[2] con la integración de un sistema de reconocimiento de voz llamado *Speech Pearl*. Este sistema se implementó en la plataforma

[2]Más información sobre el producto *Phillips Speecc Proccssing* se puede encontrar en http://www.nortelnetworks.com/corporate/news/newsreleases/2000c/09_26_0000389_philips_sp.html

OSCAR (Cómputo de Señal Abierta y Recursos de Análisis) para el comercio electrónico, que permite el uso de asistencias de directorio raíz con voz y aplicaciones de atención al cliente.

El desarrollo de esta tesis se considera un trabajo multidisciplinario, ya que involucra diversos aspectos en áreas como la linguística, procesamiento y análisis de señales, acústica, fonética, probabilidad y ciencias computacionales.

1.5 Estrategia de solución

La estrategia de solución se divide en cuatro partes:

A) Crear un corpus de voz

B) Crear modelos acústicos

C) Integrar los modelos acústicos

D) Evaluar los modelos acústicos.

1.5.1 A) Crear un corpus de voz

Un *corpus* es una colección de datos linguísticos (textos escritos o grabaciones de voz) que pueden ser usados como punto de partida en descripción linguística o en la verificación de hipótesis sobre el lenguaje [Crystal, 1991].

Para desarrollar un *corpus* de voz; se propuso una metodología la cual consta de los siguientes pasos:

1. *Diseñar el contenido linguístico del corpus.* Para crear el modelo acústico del sistema de reconocimiento de voz se requiere un *corpus* que incluya muestras suficientes de las unidades linguísticas deseadas [Listerri, 1999]. Las unidades linguísticas utilizadas en este trabajo fueron los fonemas. En este trabajo se consideró como suficiente un mínimo de 50 muestras para crear cada modelo.

2. *Obtener texto en Español.* El texto se obtuvo de tres maneras: (1) diseñando y escribiendo 505 frases manualmente, (2) capturando 2,445,566 frases automáticamente a partir de la extracción de textos de Internet [Vaufreydaz, 1999], (3) seleccionando aleatoriamente 505 frases del *corpus* DIME.

3. *Obtener la transcripción fonológica del texto obtenido.* En este proceso se aplica un conjunto de reglas fonológicas al texto obtenido. Estas reglas fonológicas se crearon a partir de la relación entre letras y fonemas del Español [Uraga y Pineda, 2002].

4. *Seleccionar aleatoriamente un conjunto de 1515 frases del corpus.* Las frases se modificaron manualmente para que cumplieran con los criterios de riqueza, suficiencia y balanceo fonéticos.

5. *Grabar las frases seleccionadas.* Las frases fueron grabadas con 50 hablantes (25 hombres y 25 mujeres). Para realizar estas grabaciones se utilizó el *CSLU Tool Kit* [CSLU, 2002], el cual permite grabar; visualizar y procesar datos de voz.

6. *Generar una transcripción automática del corpus de voz.* Este proceso se realizó mediante un programa computacional incluído en la herramienta *HTK [Young, 1997].* Este proceso utiliza el método *forced alignment*, el cual determina los límites temporales de los fonemas utilizando un sistema de reconocimiento de voz a partir de los datos de voz y su transcripción ortográfica.

1.5.2 B) Crear modelos acústicos

Este paso consistió en crear un conjunto de modelos acústicos - fonéticos a partir del *corpus* de voz. Los modelos acústicos creados en este trabajo están basados en modelos ocultos de Markov. Estos modelos fueron creados con la herramienta *HTK (Hidden Markov Model Tool Kit)* desarrollada por *Entropics[Young, 1997].* El modelado de las unidades fue independiente del contexto e independiente del hablante.

1.5.3 C) Integración de los modelos acústicos en el sistema de reconocimiento de voz

El sistema de reconocimiento de voz del proyecto DIME cuenta principalmente con tres modelos: un modelo acústico, un modelo de lenguaje y un modelo de pronunciación. El modelo acústico creado en este trabajo de tesis sustituyó al modelo acústico del sistema actual.

1.5.4 D) Evaluación del sistema de reconocimiento de voz

En la última fase se evaluó el desempeño del sistema de reconocimiento de voz utilizando los nuevos modelos acústicos. El proceso de evaluación se realizó utilizando la herramienta *HTK.* El desempeño del nuevo sistema mejoró considerablemente el desempeño del sistema anterior.

1.6 Limitaciones del trabajo

El sistema de reconocimiento de voz desarrollado tiene las siguientes limitantes:

- Los modelos acústicos solo identifican las unidades fonológicas (fonemas) del Español pero no consideran efectos de co-articulación entre fonemas.
- El vocabulario que se puede utilizar para hablar con el sistema está limitado a un conjunto de 1113 palabras.
- La aplicación que utilizó el sistema de reconocimiento de voz está restringida a diálogos sobre el diseño de cocinas.
- El sistema sólo reconoce las palabras pronunciadas de manera continua; es decir, no es robusto cuando ocurren fenómenos de habla espontánea como tartamudeos, chasquidos, interjecciones, palabras incompletas, etc.

1.7 Organización de la tesis

En el capítulo 2 se presenta brevemente en que consiste un sistema de reconocimiento de voz. Se describe cada uno de los criterios que se utilizan para clasificarlos y se explican de manera muy general cada uno de los módulos del sistema (modelo acústico, modelo del lenguaje y modelo de pronunciación).

En el capítulo 3 se presentan los criterios para clasificar a un *corpus*. Se propone una metodología para crear un *corpus* de voz rico y balanceado fonéticamente con el objetivo de crear modelos acústicos para reconocimiento de voz continua e independiente del hablante.

En el capítulo 4 se describen cada uno de los módulos de un sistema de reconocimiento de voz. Se explican la creación y entrenamiento de modelos acústicos basados en modelos ocultos de Markov. Se describe la integración de los nuevos modelos acústicos en el sistema de reconocimiento de voz.

En el capítulo 5 se presenta el método para evaluar el desempeño alcanzado por los nuevos modelos acústicos en el sistema de reconocimiento de voz continua e independiente del hablante.

Finalmente, en el capítulo 6, se muestran las conclusiones y se plantea el posible trabajo a futuro.

Capítulo 2

Sistemas de reconocimiento de voz

En este capítulo se describen los antecedentes, la descripción, la clasificación, la arquitectura, las aplicaciones, las ventajas y desventajas de los sistemas de reconocimiento de voz.

2.1 Antecedentes

En este apartado se presenta de manera cronológica como se ha desarrollado la tecnología de reconocimiento de voz.

En el siglo XVIII, se inicia el desarrollo de la tecnología de lenguaje hablado; ejemplo de ello es el trabajo de Von Kempelen que consiste en un dispositivo parlante, totalmente mecánico, capaz de emitir una veintena de sonidos; que al ser manipulado con habilidad, permitía la producción de palabras intelegibles [Casacuberta, 1987].

En 1870, Alexander Graham Bell quería construir un sistema/dispositivo que hiciera el habla visible a las personas con problemas auditivos, como resultado se obtuvo el teléfono.

Desde 1880, Tihamir Nemes había pensado en desarrollar un sistema de transcripción automática que identificara secuencias de sonidos y los imprimiera en texto; pero no obtuvo financiamiento para poder desarrollarlo.

En 1910 los laboratorios *AT&T Bell* construyeron la primera máquina capaz de reconocer la pronunciación de los 10 dígitos del Inglés. Esta máquina requería extenso reajuste de la voz de una persona, pero una vez logrado, tenía un 99% de certeza. Por lo anterior, surge la esperanza de que el reconocimiento de voz sea simple y directo.

Alrededor de 1960, los investigadores reconocen que producir sistemas de reconocimiento de voz es un proceso mucho más intrincado y sutil de lo que habían anticipado. Por lo tanto, se empiezan a reducir los alcances y se enfocan a sistemas más específicos como los que se mencionan a continuación:

- *Sistemas dependientes del hablante.-* Sólo reconocían la voz de la persona con que se había entrenado el sistema.
- *Sistemas de flujo discreto del habla.-* En estos sistemas se debía realizar una pausa entre las palabras pronunciadas.
- *Sistemas de vocabulario pequeño.-* Estos sistemas contaban con un vocabulario menor o igual a 50 palabras.

Posteriormente aparecen empresas como *IBM y CMV*, que trabajaban en reconocimiento de voz continua pero no se ven resultados hasta 1970.

A principios de 1970 comienza un proyecto de reconocimiento de voz: el VIP100 de *Threshold Technology Inc.* que utilizaba un vocabulario pequeño, dependiente del locutor y reconocía palabras aisladas .

Paralelamente, en esa década, se buscaba construir un sistema de reconocimiento de voz continua; es decir, que utilizara un vocabulario extenso y que pudiera reconocer frases pronunciadas continuamente. Además, se impulsa la investigación enfocada al entendimiento del habla, por lo que los sistemas empiezan a incorporar módulos de:

- análisis léxico (conocimiento léxico)
- análisis sintáctico (estructura de palabras)
- análisis semántico (significado)
- análisis pragmático (intención)

En 1976 se crearon sistemas para el proyecto *ARPA SUR* (*Speech Understanding Research*).

Entre 1980 y 1990 surgen sistemas de vocabulario amplio, que ahora son los más comunes (ya que utilizan más de 1000 palabras), adicionalmente bajan los precios de estos sistemas [Kirschning, 2001].

A partir de 1990 hasta la fecha, la tecnología de lenguaje hablado ha tenido un impulso considerable, puesto que actualmente existen empresas importantes dedicadas al desarrollo de tecnología de lenguaje hablado, por ejemplo:

- *Sensory Circuits*: Ha desarrollado procesadores y sintetizadores para reconocimiento de voz. Los chips creados pueden ser dependientes o independientes del hablante y tienen un alto desempeño en reconocimiento de palabras del idioma Inglés [Sensory Circuits, 2002].

- *Dragon Systems*: Desarrolló un sistema que puede reconocer alrededor de 160 palabras por minuto. Su aplicación es para dictado automático del idioma Inglés [Dragon Systems, 2002].

- *Speechworks*: Desarrolló un sistema con la ayuda de la Universidad de *Boston*. Este sistema ayuda en el aprendizaje del idioma Inglés. [Sunburst, 2002]

- *Vocalis*: Ha desarrollado sistemas comerciales que permiten realizar transacciones bancarias por medio de la voz; pueden realizarse movimientos entre cuentas de cheque, depósitos, pagos, entre otros. Las aplicaciones son para el idioma Inglés. [Vocalis Co.,2002]

- *Dialogic*: Ha desarrollado sistemas que dan solución a problemas de seguridad por medio del reconocimiento de voz. El desempeño del sistema se da en tiempo real. [1][Dialogic, 2002]

- *Novell*: Desarrolló una plataforma que permite integrar aplicaciones de reconocimiento de voz para diferentes dominios [Novell, 2002].

- *Microsoft*: Ha desarrollado proyectos que buscan solucionar problemas en la interacción humano-máquina por medio del lenguaje hablado. Se trabaja principalmente en el Modelado de Lenguaje, Aprendizaje automático de la gramática, entre otras áreas [Microsoft-MST, 2002].

2.2 Reconocimiento de voz

El objetivo central en reconocimiento de voz (RV) consiste en convertir la voz de una persona en texto, es decir, obtener la secuencia de palabras pronunciadas por una persona.

El proceso de reconocimiento de voz consiste en obtener la señal de voz, a continuación se deben extraer las características acústicas esenciales. Con estas características, se realiza una búsqueda para obtener la secuencia más probable de fonemas pronunciados, a partir de los cuales se obtendrán las palabras pronunciadas (ver figura 2.1).

[1]Los sistemas de tiempo real procesan y emiten información de forma inmediata.

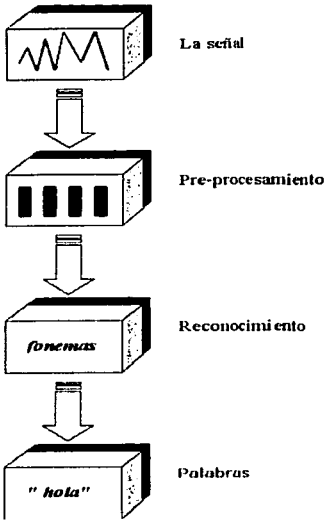


Figura 2.1: Proceso de Reconocimiento de voz.

2.3 Clasificación de sistemas de reconocimiento de voz

Los sistemas de reconocimiento de voz se pueden clasificar de acuerdo a los siguientes parámetros [Uraga, 1999]:

1. Por la dependencia del hablante
2. Por la forma de reconocimiento
3. Por las características del vocabulario

2.3.1 Dependencia del hablante

De acuerdo a la dependencia del hablante los sistemas se clasifican en:

- *Sistemas dependientes del hablante*: También se les conoce como sistemas monolocutores. Los sistemas de este tipo son entrenados con la voz de una persona; por lo que son usados específicamente para reconocer la voz de su entrenador. Por ejemplo; tenemos los sistemas de dictado automático [Dragon Systems, 2002].

- *Sistemas independientes del hablante*: Estos sistemas reconocen la voz de un grupo muy numeroso de personas. El sistema es entrenado con varios hablantes, por lo que puede trabajar con cuestiones de variabilidad en el habla, y en la mayoría de los casos, reconocer la voz de personas que no hayan participado en el entrenamiento del sistema.

2.3.2 Forma de reconocimiento

De acuerdo a la forma de reconocimiento los sistemas se clasifican en:

- *Sistemas de reconocimiento de palabras aisladas*: ésta es la forma más simple de reconocimiento. En este tipo de sistemas la pronunciación debe ser realizada con una pequeña pausa entre dos palabras consecutivas, de esta manera ninguna palabra afecta la pronunciación de otra palabra. El trabajo desarrollado por el Grupo de Tecnología del Habla ("proyecto POLYGLOT IWSR") es un ejemplo de un sistema de reconocimiento de palabras aisladas que puede reconocer 8000 palabras y requiere bajos requerimientos computacionales [Polyglot, 2002].

- *Sistemas de reconocimiento de voz continua*: Esta forma de reconocimiento implica cierta complejidad porque la voz a reconocer ha sido producida de manera natural en un determinado dominio. En estos sistemas se definen unidades linguísticas que pueden ser fonemas, sílabas, alófonos etc. Un sistema de reconocimiento de voz continua es *Speech Works*. Este sistema logra reconocer alrededor de 160 palabras por minuto [Sunburst, 2002].

- *Sistemas de reconocimiento de voz espontánea*: Estos sistemas tienen mayor nivel de complejidad; ya que esta forma de reconocimiento acepta cualquier forma de hablar y de pronunciar, en ella se pueden producir chasquidos o ruidos que pueden interferir y que aparecen en el habla natural. Es la forma de expresión oral más natural que existe.

  Aunque alcanzar un alto porcentaje en reconocimiento de voz espontánea es una tarea difícil, actualmente, en el proyecto DIME [Pineda, 2001] se utiliza un modulo de reconocimiento de voz espontánea que alcanza un 50% de reconocimiento de palabras.

2.3.3 Características del vocabulario

Las características del vocabulario varían mediante dos criterios:

1. Los sistemas pueden ser, por el número de lenguas que contienen:
   - *Monolingües*: Los sistemas con vocabulario monolingüe contienen palabras de una sola lengua. Por ejemplo tenemos el corpus DIME [Villaseñor, 2000] que sólo contiene grabaciones en Español.
   - *Multilingües*: Los sistemas multilingües involucran vocabulario en más de una lengua. Por ejemplo tenemos el proyecto CSTAR II que involucra vocabulario en Inglés, Francés, Italiano, Japonés, Alemán y Koreano [CSTAR II, 2002].
2. De acuerdo al tamaño del vocabulario, los sistemas de reconocimiento de voz se clasifican en [Gibbon en Uraga, 1999]:
   - *De vocabulario pequeño*: El vocabulario está en el rango de 1 a 100 palabras.
   - *De vocabulario mediano*: El vocabulario está en el rango de 100 a 1000 palabras.
   - *De vocabulario grande*: El vocabulario está en el rango de 1000 a 5000 palabras.
   - *De vocabulario muy grande*: El vocabulario está en el rango de 5000 palabras o más.

   En este trabajo de tesis se desarrolló un sistema de reconocimiento de voz continua, independiente del hablante, monolingüe (en español) y con un vocabulario grande (1113 palabras).

2.4 Arquitectura general de un sistema de reconocimiento de voz

En esta sección se muestra la arquitectura de un sistema de reconocimiento de voz, se describen sus distintos componentes y las funciones que realizan para convertir una señal acústica en texto.

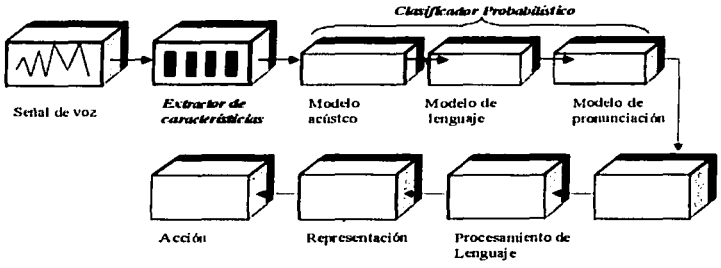


Figura 2.2: La arquitectura general de un sistema reconocimiento de voz cuenta principalmente con dos módulos: [Ahuactzin, 1999], uno que obtiene características de la señal de voz (Extractor de características) y el otro probabilístico que está integrado por un modelo acústico, un modelo de lenguaje y un modelo de pronunciación (clasificador probabilístico).

2.4.1 Extractor de características

Al momento de pronunciar una frase se genera una señal acústica. La extracción de características consiste en segmentar la señal acústica en un conjunto de vectores. Estos vectores reflejan las características acústicas de la señal de voz.

MFCC (*Mel Frecuency Cepstral Coefficients*) es una de las técnicas utilizadas para segmentar y crear los vectores de características acústicas. Al aplicarla, un número determinado de muestras se reduce de la señal de voz a un conjunto de coeficientes. Estos coeficientes representan las concentraciones de energía y anchos de frecuencia de la señal de voz.

2.4.2 Clasificador probabilístico

Este módulo conforma la estructura medular del sistema de reconocimiento de voz. Se compone de tres partes: un modelo acústico, un modelo de lenguaje y un modelo de pronunciación.

- *Modelo acústico.* Para crear un modelo acústico existen varios enfoques: redes neuronales, modelos ocultos de Markov y modelos híbridos (basado en modelos ocultos de Markov y redes neuronales). Los modelos que han tenido mejores resultados utilizan modelos ocultos de Markov y en algunos casos modelos híbridos. Los modelos acústicos proporcionan un método

para calcular la probabilidad de observar la secuencia acústica cuando el hablante pronuncie una secuencia de palabras. [Uraga, 1999]. Esta probabilidad se estima a partir de varias secuencias de palabras y de la observación de sus correspondientes secuencias acústicas.

- *Modelo de lenguaje*: Este modelo proporciona al reconocedor una fuente de conocimiento, ya que describe cómo es el lenguaje o el conjunto de palabras y las frases que puede reconocer el sistema. Este modelo incluye un léxico y una gramática.

- *Modelo de pronunciación*: Contiene todo el conjunto de palabras y su correspondiente descripción de como se pronuncia cada palabra.

En el capítulo 4 de este trabajo de tesis se describen a detalle estos tres módulos. A continuación se presentan las aplicaciones, ventajas y desventajas de los sistemas de reconocimiento de voz.

2.5 Aplicaciones

Las aplicaciones de los sistemas de reconocimiento de voz son muy variadas, por ejemplo: asistentes para llenado de formas, asistentes para dictado automático, sistemas conversacionales, control por comandos, transacciones bancarias, marcado telefónico automático, programación oral, servicios de información automática, control de máquinas y herramientas, control de calidad e inspección, manejo y clasificación automática de piezas, etc.

2.6 Ventajas y desventajas

Las ventajas de utilizar el habla como interfaz entre el hombre y la máquina dependen del dominio de aplicación. Como menciona Kirschning [Kirschning, 2001], los sistemas de reconocimiento de voz tienen, por una parte, ventajas que hacen rentable su uso:

- No se requiere capacitar a los usuarios del sistema.
- El proceso de reconocimiento es más rápido que utilizar las manos como interfaz.
- Permiten a manos y ojos permanecer libres.
- Pueden utilizarse micrófonos o en su caso se puede utilizar la vía telefónica como medio de comunicación.

Por otra parte, las desventajas más notorias en el uso de sistemas de reconocimiento de voz son:

- Las condiciones para usar el sistema deben ser iguales a las condiciones con las que se entrenó.
- El ruido puede alterar el funcionamiento del sistema.
- Es indispensable hacer uso de alguna interfaz física como el micrófono.
- En sistemas de telefonía solamente es aplicable a dominios pre-determinados.

2.7 Conclusiones

En este capítulo se describió de manera muy general la arquitectura y clasificación de un sistema de reconocimiento de voz. También se presentaron algunas aplicaciones, ventajas y desventajas de utilizar la voz como interfaz humano-máquina.

En el siguiente capítulo se describe la metodología y los criterios necesarios para obtener un corpus de voz. Este recurso lingüístico (corpus de voz) funciona como base empírica para entrenar y evaluar modelos acústicos para sistemas de reconocimiento de voz.

Capítulo 3

Corpus de voz para crear modelos acústicos

Como se mencionó en el capítulo anterior, un sistema de reconocimiento de voz está compuesto por tres partes: un modelo acústico, un modelo de lenguaje y un diccionario de pronunciación.

En este trabajo de tesis se pretende crear un conjunto de modelos acústicos para el Español hablado en México, por lo que es necesario utilizar un corpus de voz con características lingüísticas específicas.

En éste capítulo se describe qué es un corpus de voz y sus clasificaciones. Además se propone una metodología para diseñar el contenido linguístico de un corpus tomando en cuenta las ventajas y desventajas de modelar diferentes unidades lingüísticas para crear modelos acústicos.

También se describe el proceso de selección, grabación y transcripción de un corpus con el objetivo de obtener un corpus de voz rico y balanceado fonéticamente.

3.1 Definición de corpus

A continuación se presentan distintas definiciones de corpus:

Un *corpus* es un conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que puede servir de base a una investigación [RAE, 1992].

Un *corpus* es una colección de piezas de lenguaje que son seleccionadas y ordenadas de acuerdo a ciertos criterios lingüísticos para que sean utilizadas como muestras del lenguaje [Sinclair en Listerri, 1999].

Un *corpus* es una colección de datos lingüísticos, textos o grabaciones de voz,

que pueden ser usados como punto de referencia para una descripción lingüística o para la comprobación de una hipótesis del lenguaje [Crystal, 1991].

Los mayoría de los sistemas de reconocimiento de voz están basados en métodos estadísticos que se construyen a partir de la recolección de datos reales del lenguaje (*corpus* de voz).

Uno de los principales problemas que involucra el desarrollo de sistemas de reconocimiento de voz es obtener recursos lingüísticos apropiados para crear modelos acústicos.

El desempeño de los modelos acústicos está en relación directa con la cantidad y tipo de información acústica con la que se entrenan tales modelos. En teoría; a mayor cantidad de datos de lenguaje recolectados para entrenar modelos acústicos, mejor será el desempeño del sistema de reconocimiento de voz. Sin embargo; el lenguaje hablado es un universo de estudio demasiado amplio y es prácticamente imposible definir en qué consiste recolectar una muestra representativa de lenguaje para entrenar y evaluar dichos modelos.

Actualmente, se hace uso de la informática ya que facilita la explotación de grandes cantidades de datos. Por esta razón más que hablar de *corpus* debería hablarse de *corpus informatizados;* asimismo, Listerri define *corpus* informático a una cierta cantidad de datos de lenguaje codificados y estandarizados de manera homogénea que fungen como muestras de lenguaje [Listerri, 1999].

3.2 Clasificación de corpus

A continuación se presentan distintas clases de corpus:

- Por niveles
- Textual u oral
- Según el porcentaje y distribución de diferentes tipos de texto
- Según la especificidad de los textos

3.2.1 Por niveles

La clasificación por niveles se divide en:

- *Corpus*: Es un conjunto homogéneo de muestras de la lengua ya sean: orales, escritas, literales, coloquiales, etc., el cual se toma como modelo

de un estado o nivel de lengua predeterminada [Listerri, 1999]. Por ejemplo; el corpus recolectado de Internet por Vaufreydaz del idioma Francés [Vaufreydaz, 1999].

- *Subcorpus*: Suele ser una selección estática de textos derivada de un *corpus* general y complejo, el cual está dividido en grupos de muestras textuales más específicas; pero también puede ser una selección dinámica de textos de un *corpus* en crecimiento [Listerri, 1999]. Por ejemplo; el subcorpus seleccionado por la metodología propuesta por Vaufreydaz (1999) para obtener un *corpus* rico y balanceado fonéticamente.

3.2.2 Textual u oral

Por lo regular, un *corpus* se encuentra en dos formas: textual y oral

- *Corpus textual*: Está constituido por muestras de lenguaje escrito. Por ejemplo; las novelas, los anuncios del periódico, las colecciones por antologías o solo poemas, etc. Estos textos pueden ser obtenidos de forma manual o automática [Listerri, 1999].
  - La forma manual es diseñada por los humanos. Por ejemplo: Escribir un conjunto de oraciones.
  - La forma semiautomática se obtiene escribiendo el texto y utilizando herramientas computacionales que facilitan el procesamiento de texto. Por ejemplo: el *corpus* obtenido para este trabajo de tesis se obtuvo mediante dos formas: Automatica (extrayendo texto de Internet) y manual (escribiendo frases).
  - La forma automática consiste en extraer colecciones digitales mediante herramientas computacionales. Por ejemplo: la recopilación de colecciones digitales en Internet.
- *Corpus oral*: Se compone básicamente de muestras de lenguaje hablado que se obtienen mediante grabaciones. Pueden considerarse transcripciones ortográficas del habla.

  Para el desarrollo de tecnologías del habla, difícilmente se concibe un *corpus* que no vaya acompañado de su correspondiente registro sonoro. El *corpus* DIME, por ejemplo, es un *corpus* grabado en Español utilizado en una aplicación de diseño de cocinas[Villaseñor, 2000].

3.2.3 Según la distribución fonológica de los diferentes tipos de texto

Esta clasificación se refiere a criterios de proporción, distribución y variedad fonológica en el contenido linguístico del corpus:

- *Corpus rico fonéticamente*: Es aquel que contiene a la gran variedad de eventos fonéticos del lenguaje en su transcripción. Es decir, sus grabaciones contienen la pronunciación de todos (o casi todos) los fonemas en todos sus posibles contextos. Por ejemplo: el *corpus* creado por Vlasta Radová para el idioma checo; el cual ha sido de gran utilidad para modelar acústicamente la lengua checa, ha obtenido mejores resultados en reconocimiento de voz; en comparación a trabajos realizados con *corpus* que no contienen toda la variedad fonética. [Radová, 1999]

- *Corpus balanceado fonéticamente*: Es aquel que busca obtener eventos fonéticos en una frecuencia aproximada en el habla natural; por ello, podemos suponer que un evento fonético ocurre un número de veces mayor o menor que otro. Por ejemplo: en el Español es mayor el empleo de vocales que de consonantes a diferencia de otras lenguas como el Alemán o el Mandarín [Radová, 1999].

  El *corpus* diseñado para el idioma Checo por Radová, fue obtenido mediante un proceso de balanceo fonético para obtener una frecuencia de fonemas en proporción aproximada al habla [Radová, 1999].

3.2.4 Según la especificidad de los textos

La siguiente clasificación se genera a partir del tipo de texto que contiene un corpus:

- *Corpus general*: Este tipo de *corpus* pretende reflejar la lengua común en su ámbito más amplio, se interesa por obtener distintas clases de textos. Este tipo de *corpus* es útil para describir la lengua común de una colectividad, es decir, el lenguaje que utilizan los hablantes en situaciones comunicativas normales [Listerri, 1999]. Por ejemplo: el *corpus* diseñado para este trabajo de tesis es general; ya que es para el Español hablado en México.

- *Corpus multilingües*: Este tipo de corpus contiene textos de diferentes lenguas sin que éstos sean, necesariamente, traducciones unos de otros y sin compartir criterios de selección [Sinclair en Listerri, 1999]. Por ejemplo: el *corpus* ECI/MCI (*European Corpus Initiative*) contiene una colección de textos en Inglés, Alemán, Francés y Español (alrededor de 5 millones de palabras).

- *Corpus especializado*: Se opone al corpus general. El *corpus* especializado recoge textos que pueden aportar datos para descripción de un tipo particular de lengua. Un ejemplo de este tipo de *corpus* es una colección de textos poéticos.

- *Corpus periódico o cronológico*: Este *corpus* contiene textos de una época determinada. Por ejemplo; la colección de textos de novelas que pertenecen únicamente al siglo XIX [Listerri, 1999].

En este trabajo de tesis se diseñó un *corpus* de voz con características lingüísticas específicas del Español hablado en México con el objetivo de obtener datos empíricos sobre dicha lengua. Se pretende además que el *corpus* funcione como base para el entrenamiento de modelos acústicos aplicados a reconocimiento de voz continua.

En la siguiente sección, se propone una metodología para obtener un corpus de voz rico y balanceado fonéticamente.

3.3 Metodología para crear un corpus de voz

En esta sección se describe la metodología utilizada en este trabajo de tesis para obtener un *corpus* de voz rico y balanceado fonéticamente. Los pasos que se siguieron son los siguientes:

- Diseño del contenido linguístico del *corpus*
- Recolección de texto
- Transcripción automática de texto
- Selección de frases
- Grabación de frases
- Transcripción automática del *corpus* de voz

Cabe señalar que los *corpus* que existen para el Español de México han sido generados sin diseño previo; por esta razón, el contenido linguístico de estos *corpus* puede resultar insuficiente para crear, entrenar y evaluar modelos acústicos.

3.3.1 Diseño del contenido lingüístico del corpus

Para entrenar y evaluar sistemas de reconocimiento se requiere un *corpus* de voz grabado con las unidades lingüísticas que se quieren reconocer [Listerri, 1999]. Las unidades lingüísticas pueden ser palabras, sílabas, trifonemas, bifonemas, y fonemas, entre otros, etc.

Tomando en cuenta que el *corpus* será utilizado para entrenar modelos acústicos en reconocimiento de voz continua e independiente del hablante, es conveniente mencionar las ventajas y desventajas que presenta cada tipo de unidad lingüística en su modelado; ya que esto permitirá definir una unidad lingüística conveniente.

- Modelado acústico a nivel de palabras
  - *Ventaja*: El modelado lingüístico a nivel de palabras permite capturar automáticamente los efectos de coarticulación en la palabra. Por ejemplo, al pronunciarse la palabra *investigador*, que se pronuncia como */imbestigador/*, se da un efecto de co-articulación entre los fonemas /n/ y /b/ que altera la pronunciación de la palabra. Al tener un modelo acústico para cada palabra, la probabilidad de identificar la palabra correcta es mayor.
  - *Desventaja*: Se obtiene un modelo acústico para cada palabra del vocabulario del *corpus*, entonces el sistema requiere de mucha memoria para dominios de vocabulario amplio.
- Modelado acústico a nivel de sílabas
  - *Ventaja*: El modelado sílabico es más específico que a nivel de palabra; además, captura efectos de coarticulación dentro de la sílaba. Por ejemplo: la palabra *investigador* está compuesta por 5 sílabas *in-ves-ti-ga-dor*, donde para cada sílaba se tiene un modelo acústico.
  - *Desventaja*: Son demasiadas sílabas, para el Español se tienen más de 2000. Además no captura efectos de co-articulación entre sílabas. [Kirschning, 2001]
- Modelado acústico a nivel de trifonemas
  - *Ventaja*: El modelado a nivel de trifonemas, permite capturar fenómenos de co-articulación en diferentes contextos.
  - *Desventaja*: Se requiere un gran corpus de voz para entrenar modelos a nivel de trifonema; además, es difícil encontrar todos y cada uno de los trifonemas en los datos de entrenamiento.
- Modelado acústico a nivel de bifonemas

- *Ventaja*: Si se modelan bifonemas, se van a capturar efectos de co-articulación entre fonemas en su contexto derecho e izquierdo. Por ejemplo; el fonema /*a*/, en la palabra *casa* se co-articula en su contexto izquierdo con el fonema /*k*/, y con el fonema /*s*/ en su contexto derecho, entonces, los fenómenos acústicos entre fonema y fonema son modelados en ambos contextos.
- *Desventaja*: Son demasiadas unidades (bifonemas). Además, no modela efectos de co-articulación entre bifonemas.

- Modelado acústico a nivel de fonemas
  - *Ventaja*: Es un conjunto finito y resulta más sencillo encontrar muestras de todo el inventario de fonemas; además, resulta más económico generar un modelo para cada fonema que un modelo para cada sílaba, trifonema o bifonema.
  - *Desventaja*: El modelado a nivel fonémico no permite capturar efectos de co-articulación entre fonemas y esto afecta el desempeño del sistema debido a que; los fonemas son afectados por su contexto.

En este trabajo se decidió crear modelos acústicos independientes del contexto; es decir, no se considera cómo se afecta la pronunciación entre sonidos por efectos de coarticulación; y la unidad escogida para modelar la voz acústicamente es el fonema; ya que su modelado es independiente del dominio de aplicación y es menos costoso generar un modelo para cada fonema que generar un modelo para cada palabra, sílaba, bifonema, trifonema, etc.

Los fonemas son los sonidos que conforman una lengua; además forman las unidades de estudio de la fonología; es decir, la fonología es la disciplina lingüística que se encarga del estudio y clasificación de los sonidos, basándose en las características de articulación y en la distribución en la cadena sonora del habla.

Los fonemas se representan entre dos diagonales / /. Por ejemplo: el fonema /s/ representa el sonido producido al pronunciar las letras *s y z* en las palabras *casa* y *caza*, respectivamente.

La fonología distingue 22 fonemas para el Español hablado en México, clasificados en 5 vocales y 17 consonantes [Munguía, 1998] (ver tabla 3.1).

Para entrenar modelos acústicos, es necesario obtener muestras suficientes de cada unidad que se quiera modelar. Sin embargo, en el habla natural, los fonemas se pronuncian en proporciones diferentes e intuitivamente podemos suponer que las vocales ocurren con mayor frecuencia que las consonantes y además, entre las consonantes existen algunas que ocurren un número mayor de veces que otras.

Cuando se trabaja con lenguaje escrito, es prácticamente imposible delimitar nuestro universo de estudio; para esto tendríamos que conseguir todos y cada

/a/	/n/
/b/	/ñ/
/ch/	/o/
/d/	/p/
/e/	/r/
/f/	/rr/
/g/	/s/
/i/	/t/
/k/	/u/
/l/	/x/
/m/	/y/

Tabla 3.1: Inventario de fonemas del Español de México.

uno de los escritos que existen para el Español. Situación semejante se presenta cuando se trabaja con muestras de lenguaje hablado, si se quisiera recolectar una muestra del lenguaje que incluyera a todas las formas de pronunciación de las unidades lingüísticas para el Español, se tendría que contar con las grabaciones de todas y cada una de las personas que las pronuncian, lo cual resulta imposible.

Sin embargo, el tamaño de la muestra podría estimarse por medio de algún método estadístico, siempre y cuando toda variable (unidad lingüística), tenga una probabilidad de aparición definida en el universo, pero a su vez no se pueden estimar proporciones cuando el universo de estudio es demasiado grande y se modifica continuamente. Por lo tanto, el tamaño de la muestra requerida puede oscilar y por ello ocasionar que sea difícil obtener estadísticamente la frecuencia de aparición de fonemas, sílabas, palabras, etc., en el lenguaje hablado.

Sin embargo, hay algunas publicaciones en las que se reporta una estimación de las frecuencias de aparición de fonemas (ver tabla 3.2).

Es conveniente entrenar a los modelos acústicos con una frecuencia similar al orden de aparición en el habla de una lengua en particular[Radová, 1999]. Una consecuencia favorable de adoptar este criterio radica en conseguir una frecuencia aproximada para cada unidad lingüística; ya que a partir de dicha frecuencia se calcula la probabilidad de cada unidad linguística. Esta probabilidad se incluye en los modelos acústicos y es determinante al momento de distinguir entre diferentes unidades que son similares acústicamente (como */n/ y /ñ/*). De otra forma, si se asigna a todos los fonemas una probabilidad uniforme, el sistema podría confundir fácilmente a diferentes fonemas como */ n / y / ñ /* y reconocer la palabra *caña* en lugar de *cana* y viceversa.

Debido a que todos los modelos acústicos utilizados para reconocimiento de voz se crean con métodos estadísticos que calculan la probabilidad de ocurrencia de una secuencia de sonidos (ver siguiente capítulo), es necesario crear un corpus que

Fonema	Frecuencia [Delattre, 1965]	Frecuencia [Listerri, 1999]	Frecuencia [Rojo, 1991]
/a/	12.97	13.43	13.40
/b/	2.86	2.92	2.66
/ch/	0.32	0.40	0.28
/d/	5.21	3.96	4.79
/e/	14.00	13.72	13.51
/f/	0.52	0.51	0.68
/g/	0.71	0.90	0.95
/i/	4.47	6.89	7.5
/k/	4.65	4.04	3.98
/l/	3.85	4.25	5.08
/ll/	0.47	0.54	0.38
/m/	3.74	3.63	3.09
/n/	7.03	7.48	6.99
/ñ/	0.30	0.46	incluída en la /n/
/o/	9.21	10.37	9.57
/p/	2.30	2.6	2.66
/r/	6.24	0.40	0.79
/rr/	incluída en /r/	4.25	5.67
/s/	8.37	8.28	7.58
/t/	4.75	4.63	4.48
/u/	3.06	3.33	3.16
/x/	0.37	0.63	0.73
/y/	0.47	0.19	0.22
/z/	1.42	1.53	1.68

Tabla 3.2: Frecuencia aproximada de fonemas del Español.

contenga, para cada fonema, una frecuencia aproximada a su orden de aparición en el lenguaje natural.

Para obtener muestras suficientes de fonemas es necesario recolectar la mayor cantidad de texto posible. A continuación se describe dicho proceso.

3.3.2 Recolección de texto

La recolección de texto se logró mediante dos formas: Automática y Manual.

El proceso de extracción automática obtuvo texto a partir de dos fuentes:

- De internet
- Del corpus DIME

Extracción automática de textos de Internet

Los documentos se extrajeron de internet por medio de un programa computacional (ver apéndice A). Los documentos se obtuvieron en un formato html (*hypertext markup languague*); por lo tanto, la información no consistía de texto puro, puesto que contenía caracteres especiales que etiquetan el documento para darle atributos específicos. Para lograr obtener texto sin caracteres especiales (abreviaturas, números, etc.) se realizó el siguiente proceso:

1. Extraer un conjunto de documentos en formato HTML. (ver figura 3.1)

```
</html>
</head>pagina del periodico La Jornada<head>
</title>Antologia de poetas Mexicanos UNAM<title>
<body bgcolor=#ffffff text=#000000>
México, D.F. a 13 de marzo de 1999.
UNAM, semillero de poetas de corte fino, mención del Colegio de Poetas
Hispanoamericanos.
Una situación que engrandece la dignidad humana es la libertad de
pensamiento ejercida bajo la propia voluntad del individuo.
</body>
.....
```

Figura 3.1: Estructura de un documento en formato HTML.

2. Convertir el formato HTML a Texto (ver figura 3.2). Esta tarea se realizó automaticamente (ver apéndice B).

```
pagina del periodico La Jornada Antologia de poetas Mexicanos
UNAM
México, D.F. a 13 de marzo de 1999.
UNAM, semillero de poetas de corte fino, mención del Colegio de
Poetas Hispanoamericanos.
Una situación que engrandece la dignidad humana es la libertad de
pensamiento ejercida bajo la propia voluntad del individuo.
.....
.....
```

Figura 3.2: Documento en formato de texto.

3. Transcribir las cifras númericas a palabras. La tarea consistió en transcribir automáticamente (ver apéndice C) los números de las fechas a las palabras correspondientes. (ver figura 3.3).

1. pagina del periodico La Jornada .
2. Antologia de poetas Mexicanos UNAM.
3. México, D.F. a trece de marzo de mil novecientos noventa y nueve.
4. Semillero de poetas de corte fino, mención del Colegio de Poetas Hispanoamericanos.
.....
.....
.....

Figura 3.3: Los números se transcribieron a texto.

1. pagina del periodico La Jornada .
2. Antologia de poetas Mexicanos Universidad Nacional Autonoma de México.
3. México, Distrito Federal a trece de marzo de mil novecientos noventa y nueve.
4. Semillero de poetas de corte fino, mención del Colegio de Poetas Hispanoamericanos.
.....
.....
.....

Figura 3.4: Transcripción de acrónimos a su significado no abreviado.

4. Transcribir las siglas (acrónimos) a su significado no abreviado. La tarea consiste en generar automáticamente (ver apéndice C) un diccionario de acrónimos que aparecen en el corpus de texto. Los acrónimos se sustituyen por las palabras que representan. (Ver figura 3.4)

En total, se obtuvieron 2,445,566 frases y un total de 13,118,612 palabras. A continuación se muestran 10 frases obtenidas de Internet.

1. Exhibirán por primera vez obras inconclusas de Diego Rivera.
2. Decisiva, la participación de cinco actrices en el proyecto.
3. Hablar de lo femenino permite incluir tolerancia y otredad.
4. Desconoce la oposición el veto de Montiel a la reforma electoral.
5. Le pedirán que publique los cambios en la Gaceta de Gobierno.
6. Incluye también exhorto de búsqueda y captura internacional.
7. Los universitarios no pierden la esperanza de colarse a la liguilla.
8. Ninguna respuesta a la carta que le envió Gobernación, señala.

9. El virtuosismo siempre debe servir a la música y a la escencia de la vida.
10. La Cruz Roja de Alemania apoyará la reconstrucción.

Extracción automática de frases del corpus DIME

Algunas frases fueron extraídas del corpus DIME [Villaseñor, 2000]. Este corpus contiene 3177 frases y un total de 18065 palabras. Las siguientes frases pertenecen al corpus DIME.

1. bueno no recuerdo bien qué era eso
2. qué objeto es éste
3. sí
4. okey
5. entonces quisiera mover esta alacena al lado de la estufa que está aquí
6. bueno no tal cual en esa posición
7. tendríamos que girarlo
8. híjole tal vez está un poco separado de la estufa
9. podrías juntarlo un poco más
10. bueno parece que está

Recolección manual

La forma manual consiste en escribir texto. En este trabajo de tesis se escribieron frases que incluyen las combinaciones fonéticas del Español. Este proceso se describe a continuación:

1. Crear una lista de combinaciones entre fonemas del Español Mexicano. Por ejemplo, la tabla 3.3 muestra algunas de las combinaciones encontradas para el fonema /*b*/ .
2. Escribir palabras que contengan una o más combinaciones entre fonemas.
3. Escribir oraciones que contengan combinaciones entre fonemas. Por ejemplo; para el fonema /*b*/:
   - Un ca*mb*io es una alteración.

Contexto Izquierdo	Fonema	Contexto Derecho	Palabra
m	b		ca*mb*io
	b	r	a*br*azados
	b	l	ha*bl*o
	b	vocal	*b*arco

Tabla 3.3: El fonema /*b*/ y algunas de sus posibles combinaciones fonéticas.

- Amanecimos a*br*azados en medio de la playa.
- Yo ha*bl*o de lo que me parece real.
- Mi padre viajaba en aquel *b*arco.

4. Verificar que todas las combinaciones entre fonemas aparezcan en el conjunto de oraciones.

En total se escribieron 505 frases y 4458 palabras. Las 433 combinaciones entre fonemas se incluyeron en las palabras que conforman las 505 frases. A continuación se presentan 10 frases del corpus diseñado y escrito manualmente.

1. El gozne se ha vencido
2. El abyecto anciano caminaba por las calles
3. La abyección ha invadido esta colonia
4. Tenemos aquí como ejemplo un ángulo adyacente
5. Debemos coadyuvar al entrenamiento del sistema
6. Tuvo que enyugar para obtener el perdón
7. La pihuela sujeta perfectamente el erguido cóndor
8. Acomoden el papel a manera de resma
9. La loa resultó muy efusiva
10. La vida es el mejor ejemplo de axioma

3.3.3 Transcripción de Texto a Fonemas

Una transcripción representa elementos fonéticos, fonológicos, léxicos o morfológicos de una lengua o dialecto mediante un sistema de escritura [RAE, 1992].

La transcripción fonológica consistió en transcribir los fonemas del corpus. Este tipo de transcripción representa la pronunciación de una expresión a con fonemas; es decir, los hablantes de cualquier lengua utilizan sonidos articulados para formar palabras, que al escribirse se representan por letras.

El proceso de transcripción se realizó mediante un método que aplica reglas fonológicas a una secuencia de palabras para convertirlas de texto a fonemas.

Según la Real Academia de la Lengua Española [Munguía, 1998] se han localizado 26 letras que integran el sistema de escritura y 23 fonemas que integran el sistema fonológico del Español.

La relación entre letras y fonemas en el Español no es uno a uno, debido a que algunas de las letras pueden ser pronunciadas en más de una forma dependiendo del contexto de aparición de una letra en una palabra, además, diferentes fonemas pueden ser representados por la misma letra. Por lo que, las reglas fonológicas aplicadas para obtener la transcripción de texto a fonemas se derivan de su relación entre sí.

A continuación se explican las reglas para el sistema fonológico del Español. [Uraga y Pineda, 2002]

1. Una letra representa un fonema. Por ejemplo, *t* representa solo un fonema */t/*.

2. Una misma letra puede representar diferentes fonemas en diferentes contextos. Por ejemplo, la letra *"x"* puede representar */j/*, */s/* o /ch/ como ocurre en las palabras '*México*', '*excepción*' y '*mixiotle*'.

3. Una letra puede representar una secuencia de dos fonemas diferentes. Por ejemplo, la palabra '*excelente*' es pronunciada como la secuencia */k/ /s/*.

4. Una secuencia de letras puede representar un fonema simple. Por ejemplo, *"gu"* antes de *"e"* ó *"i"* representa el fonema */g/* como sucede en la palabra '*águila*'.

5. Diferentes letras puede representar el mismo fonema. Por ejemplo, en las palabras '*zumo*', '*sala*', '*excepción*' y '*cero*', todas las letras *z*, *s*, *x* y *c* (antes de e ó i) son pronunciadas como */s/*.

6. Algunas letras no representan ningun fonema. Una *h* muda es omitida en la pronunciación si la letra *c* no aparece antes de la *h*, como sucede en las palabras '*hola*' o '*zanahoria*'.

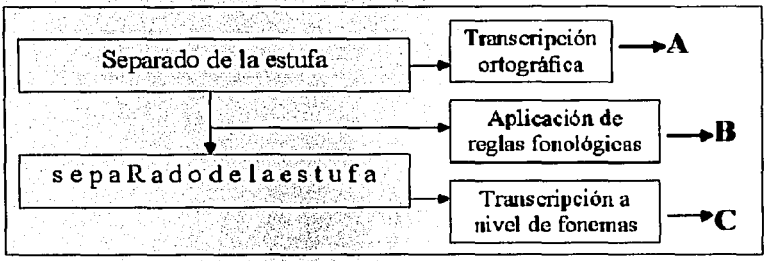


Figura 3.5: Proceso de transcripción fonológica.

7. La misma letra, en el mismo contexto, puede representarse por diferentes fonemas en diferentes palabras dependiendo de las fronteras silábicas. La palabra '*sobra*', por ejemplo, está formada por las sílabas *so* + *bra* y *r* es pronunciado con */r/*, mientras la palabra '*subraya*' está formada por las sílabas *sub* + *ra* + *ya* (b y r en diferentes sílabas) y r es pronunciado con */rr*

8. La aparición de algunos caracteres especiales como la dieresis () y el acento () determinan si una letra está pronunciada en una forma u otra. La dieresis en la vocal *u* se usa para indicar la pronunciación de *u* entre *g* y *e* o *i*. Por ejemplo, en las palabras 'antigüedad', 'pingüino'; el acento en las vocales: *a,e,i,o,u* puede determinar las fronteras silábicas de las palabras. La palabra *río* es pronunciada con dos sílabas *rí-o* donde la primera sílaba está acentuada.

9. La misma palabra está pronunciada en diferentes formas por diferentes personas. Por ejemplo. '*área*' está pronunciada con dos sílabas *á-rea* o en tres sílabas *á-re-a*. La palabra '*cae*' está pronunciada con una sílaba cai, o por dos sílabas *ca-e*.

El proceso de aplicación de reglas fonológicas se realizó automáticamente mediante un programa computacional. (ver apéndice D).

La figura 3.5 representa el proceso para transcribir una frase de nivel ortográfico a nivel fonético. En el cuadro A, se tiene una frase a nivel ortográfico; en el cuadro B se ubica el proceso de aplicación de reglas fonológicas a la frase del cuadro A y finalmente; en el cuadro C se tiene la transcripción fonológica.

3.3.4 Selección de frases

Como se ha mencionado ya, obtener un corpus especial para entrenar modelos acústicos es un proceso que involucra seleccionar frases, que contengan variedad linguística, en proporciones aproximadas a su orden de aparición en el habla.

En ésta sección se describe el proceso de selección: Dado un conjunto de frases y su transcripción a nivel de fonemas, y dada una distribución estadística de los fonemas (ver tabla 3.2), el problema consiste en seleccionar un subconjunto de frases con las siguientes características:

1. *Riqueza fonética*: consiste en que cada frase debe ser fonológicamente variada.
2. *Suficiencia fonética*: las frases deben tener una frecuencia mínima de 50 muestras para cada unidad linguística.
3. *Balanceo fonético*: Las frases deben contener una distribución estadística proporcional a una distribución dada.

La solución del problema se realizó mediante los siguientes procesos:

- Proceso de ordenamiento del corpus.
- Proceso de selección de frases.

El Proceso de ordenamiento del corpus consistió en:

1. Calcular el número de eventos fonéticos en el corpus (ver apéndice D).
2. Calcular el número de veces que cada evento fonético ocurre en cada frase.
3. Crear una tabla con cada frase del corpus y su número de eventos fonéticos distintos.
4. Ordenar las frases de acuerdo a su número de eventos fonéticos distintos (de manera ascendente).
5. Definir que el número mínimo de muestras para cada evento fonético es 50.

El proceso de selección de frases consistió en:

1. Crear una lista de eventos fonéticos ordenada (de manera ascendente) de acuerdo a su número de ocurrencias.

2. Hasta que el corpus tenga todos los eventos fonéticos y un número de muestras mayor a 50[1]

   (a) Para cada evento fonético en la lista
   Si el evento fonético no está en el corpus seleccionado o si el evento fonético ocurre un número de veces menor a 50 muestras
   entonces se debe escoger una frase del corpus que tenga tal evento fonético e incluirla en el conjunto de frases.

3. Hasta que la frecuencia de cada evento fonético sea igual a dicha frecuencia ±10% de su valor:

   (a) Seleccionar aleatoriamente una frase del corpus.

Se seleccionaron 1515 frases (14204 palabras y 86010 fonemas). La tabla 3.3 muestra la frecuencia de aparición de cada fonema y la proporción de cada unidad en el corpus seleccionado.

La tabla 3.4 compara la distribución fonética del corpus seleccionado con las proporciones obtenidas por otros trabajos de investigación [2]. El proceso de selección de frases permite obtener el texto que será grabado para obtener un corpus de voz.

Cabe resaltar que los tres fonemas que más aparecen son las vocales /e, a, o/. éstas cubren un 40% de la frecuencia total de todos los fonemas del corpus.

3.3.5 Grabación de frases

Con la finalidad de crear un corpus de voz, se realizaron grabaciones a 50 personas; de ellas 25 son mujeres y 25 son hombres, cuya lengua materna es el español de México, 42 hablantes nacieron en la capital del país (distrito federal) y el resto nació en el interior de la República Mexicana, de los cuales 2 son de Chiapas, 2 de Mexicali, 2 de Veracruz, 1 de Guadalajara y 1 de Puebla. De estos, 47 radican en la capital del país y 3 en el Estado de México.

[1] Este número se puede definir arbitrariamente pero en la literatura se recomiendan al menos 50 muestras para cada evento fonético para entrenar modelos acústicos.

[2] Aunque, el análisis estadístico fue desarrollado con base en un corpus de texto cuyo estilo literario no contiene todos y cada una de las formas de expresión coloquial del Español, tales como revistas, novelas, etc., , es cierto que, excepto para dominios específicos, la mayoría de las palabras aparecen en los diferentes dominios. Este trabajo de tesis ha escogido modelar unidades acústicas a nivel de fonema, debido a que las 22 unidades localizadas aparecen en todos los dominios independientemente de su estilo literario.

Fonema	Frecuencia por Fonema	Porcentaje en el corpus
/a/	8904	13.50
/b/	1712	2.59
/ch/	133	0.20
/d/	3137	4.75
/e/	9400	14.25
/f/	658	0.99
/g/	666	1.01
/i/	4492	6.81
/k/	2705	4.10
/l/	3618	5.48
/m/	1755	2.66
/n/	4344	6.58
/ñ/	118	0.17
/o/	6095	9.24
/p/	1705	2.58
/r/	4116	6.24
/rr/	462	0.70
/s/	6001	9.10
/t/	3036	4.60
/u/	1424	2.16
/x/	510	0.77
/y/	145	0.22

Tabla 3.4: Frecuencia de aparición de fonemas en el corpus seleccionado.

Fonema	Porcentaje [Delattre, 1965]	Porcentaje [Listerri, 1999]	Porcentaje [Rojo, 1991]	Porcentaje corpus
/a/	12.97	13.43	13.40	13.50
/b/	2.86	2.92	2.66	2.59
/ch/	0.32	0.40	0.28	0.20
/d/	5.21	3.96	4.79	4.75
/e/	14.00	13.72	13.51	14.25
/f/	0.52	0.51	0.68	0.99
/g/	0.71	0.90	0.95	1.01
/i/	4.47	6.89	7.5	6.81
/k/	4.65	4.04	3.98	4.10
/l/	3.85	4.25	5.08	5.48
/ll/	0.47	0.54	0.38	No aplica
/m/	3.74	3.63	3.09	2.66
/n/	7.03	7.48	6.99	6.58
/ñ/	0.30	0.46	incluída en la /n/	0.17
/o/	9.21	10.37	9.57	9.24
/p/	2.30	2.6	2.66	2.58
/r/	6.24	0.40	0.79	6.24
/rr/	incluída en /r/	4.25	5.67	0.70
/s/	8.37	8.28	7.58	9.10
/t/	4.75	4.63	4.48	4.60
/u/	3.06	3.33	3.16	2.16
/x/	0.37	0.63	0.73	0.77
/y/	0.47	0.19	0.22	0.22
/z/	1.42	1.53	1.68	No aplica

Tabla 3.5: Porcentaje de fonemas en el Español

Las grabaciones consistieron en leer de manera natural el conjunto de frases. La lectura de cada frase no permitía interrupciones, estornudos, tartamudeos, etc.

El trabajo se distribuyó de la siguiente manera: a 47 hablantes se le asignaron 30 frases y sólo a tres 35 frases; es decir, se grabaron un total de 1515 frases. Para cada frase se obtuvo un archivo de audio (*wav*). En cada archivo se verificaron propiedades de audio (ruido, volumen, etc.).

Las grabaciones se realizaron por medio de una computadora, un micrófono y un par de bocinas. La herramienta utilizada fue el *CSLU Tool Kit* [CSLU, 2002].

Las grabaciones se realizaron en un ambiente normal de oficina, se trató de grabar continuamente para obtener una pronunciación sin ruidos, chasquidos, interjecciones, etc.

3.3.6 Transcripción automática del corpus de voz

El último paso de la metodología consistió en transcribir los datos de voz a su correspondiente transcripción fonética. Este procedimiento se realizó mediante una herramienta computacional (*HTK*); utilizando el procedimiento de *forced alignment*. Este procedimiento permite determinar límites fonéticos utilizando técnicas de reconocimiento de voz para localizar los fonemas, dado solamente el texto de lo que se dijo y la señal grabada.

El proceso que sigue *forced alignment* para la determinación de los límites es el siguiente: primero, la señal es representada en vectores de características, y segundo, el reconocedor evalúa estos vectores reconociendo a que unidad fonética corresponden, dado que se conoce el texto de la señal. De esta manera la salida del reconocedor es la identificación de las unidades fonéticas [Olivier, 1999].

El método *forced alignment* reduce el problema de búsqueda que realiza el algoritmo de reconocimiento de voz , debido a que al conocer el texto correspondiente a la señal, se restringe la búsqueda a la secuencia de fonemas conocida.

En la Figura 3.6 se describe el proceso de *forced alignment*: El inciso A) muestra los datos de voz de la frase "*La almohada está*". El inciso B) contiene la transcripción ortográfica de la frase, C) muestra los límites temporales del fonema /*m*/, D) representa la transcripción a nivel de fonemas; en este nivel cada fonema está limitado por su tiempo de duración en la frase y E) es la transcripción ortográfica.

Cabe resaltar que una vez que se procesan los datos de voz y su correspondiente transcripción ortográfica, por medio del proceso *forced alignment*, se obtiene una transcripción a nivel de fonemas con información sobre los límites temporales de cada fonema. Esta información se almacena en archivos y posteriormente, se utiliza para entrenar modelos acústicos.

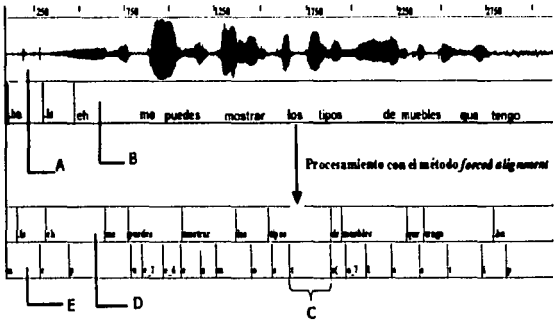


Figura 3.6: Proceso de *forced alignment*.

Una vez realizada la transcripción de los datos de voz se cuenta con los datos necesarios para crear los modelos acústicos. Este tema se desarrolla en el siguiente capítulo.

3.3.7 Conclusiones

En este capítulo se definió qué es un corpus y sus distintas clasificaciones. Se describió, la metodología utilizada para diseñar el contenido lingüístico de un corpus, mencionando las ventajas y desventajas de modelar acústicamente las diferentes unidades lingüísticas de una lengua.

Se desarrolló, el proceso para recolectar, transcribir, seleccionar y grabar texto con el objetivo de conseguir un corpus de voz rico y balanceado fonéticamente para entrenar modelos acústicos. Además, se explicó el procedimiento *forced alignment* utilizado para realizar la transcripción de voz automáticamente.

El corpus conseguido contiene datos del Español hablado en México; contiene todos y cada uno de sus fonemas con una distribución aproximada a su frecuencia de aparición en el habla.

El corpus es muy útil para el área de procesamiento de lenguaje y, más aún, en la creación y entrenamiento de modelos acústicos para reconocimiento de voz continua. En el siguiente capítulo se describen los modelos acústicos, modelos de lenguaje y modelos de pronunciación que integran a un sistema de reconocimiento de voz.

Capítulo 4

Construcción del Sistema de Reconocimiento de Voz

En esta de tesis se trabajó en el módulo de reconocimiento de voz del proyecto DIME (Diálogos Inteligentes Multimodales para el Español). Este módulo se compone de un modelo acústico, un modelo de lenguaje y un modelo de pronunciación. Debido a la complejidad en la tarea de reconocimiento de voz, se decidió crear un nuevo modelo acústico para reconocimiento de voz continua, entrenados y evaluados con un corpus de voz rico y balanceado fonéticamente. El objetivo de crear este nuevo modelo es mejorar el desempeño del sistema de reconocimiento de voz anterior. A continuación se plantea formalmente el problema de reconocimiento de voz.

4.1 Planteamiento del problema

Desde el punto de vista probabilístico, el problema de reconocimiento de voz consiste en obtener la secuencia de palabras $W = w_1, w_2, ..., w_n$, tal que:

$$W = max_w P(W|O)$$

Considerando que este problema se divide en dos partes, una relacionada con el aspecto acústico y otra con el lenguaje, éste se puede plantear formalmente con el teorema de Bayes [Charniak, 1993 en Uraga, 1999]:

$$P(W|O) = \frac{P(W)P(O|W)}{P(O)}$$

donde $P(W|O)$ representa la probabilidad *a posteriori* de haber pronunciado la secuencia de palabras W dada una secuencia de observaciones acústicas O, $P(O|W)$ corresponde a la probabilidad de observar la secuencia O cuando el hablante ha pronunciado una secuencia de palabras W, $P(W)$ se interpreta como

la probabilidad *a priori* de que el hablante pronuncie la secuencia W y $P(O)$ es la probabilidad de que O haya sido observada. Como $P(O)$ es un factor de normalización, la secuencia de W más probable se obtiene encontrando la secuencia de palabras que maximiza el producto de $P(W)$ y $P(O|W)$ [Uraga, 1999]:

$$W = \max_{w} P(W|O) = \max_{w} P(W)P(O|W) \tag{4.1}$$

donde W corresponde a la secuencia candidata con la probabilidad máxima *a posteriori* (MAP). $P(W)$ se calcula con un modelo de lenguaje, mientras que $P(O|W)$ se calcula mediante un modelo acústico. En la siguiente sección se describen los modelos de lenguaje.

4.2 Modelos de Lenguaje

El propósito de un modelo de lenguaje es estimar la probabilidad de una secuencia de n palabras $W = w_1, w_2, ..., w_n$ de un vocabulario definido. Esta probabilidad se expresa como $P(W)$. En reconocimiento de voz, $P(W)$ se interpreta como la probabilidad *a priori* de que el hablante pronuncie la secuencia W. Esta probabilidad guía la búsqueda del reconocedor entre varias opciones y es un factor que contribuye al reconocimiento final. Formalmente, $P(W)$ se puede calcular como [Charniak, 1993 en Uraga, 1999]:

$$P(W) = P(w_1, .., w_n) = P(w_1)P(w_2|w_1)P(W_3|w_1, w_2)...P(W_n|w_1, .., W_{n-1})$$

donde $P(w_i|w_1, ..., w_{i-1})$ es la probabilidad de que w_i sea pronunciada dado que la secuencia $w_1, ..., w_{i-1}$ haya sido pronunciada.

Algunos de los modelos de lenguaje que han dado buenos resultados son los modelos estadísticos n-gramas[1]: bigramas y trigramas. Los modelos se contruyen a partir de los archivos de transcripción ortográfica de un corpus y posteriormente, el texto es procesado (ver figura 4.1) para obtener la lista de monogramas, bigramas o trigramas del modelo. Finalmente, se realiza un análisis estadístico con el objetivo de estimar la probabilidad para cada *n-grama* (ver figura 4.2).

[1] n es el número de palabras

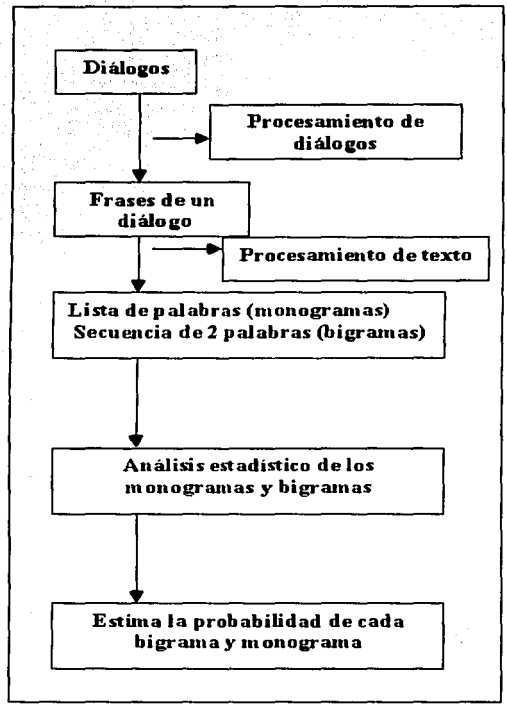
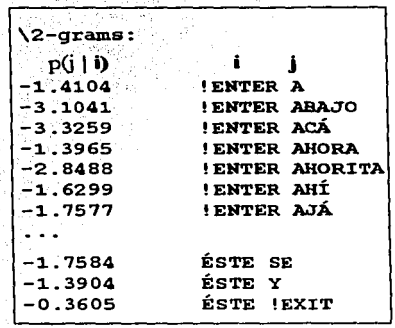


Figura 4.1: Proceso para crear bigramas.



```
\2-grams:
  p(j | i)          i     j
-1.4104          !ENTER A
-3.1041          !ENTER ABAJO
-3.3259          !ENTER ACÁ
-1.3965          !ENTER AHORA
-2.8488          !ENTER AHORITA
-1.6299          !ENTER AHÍ
-1.7577          !ENTER AJÁ
...
-1.7584          ÉSTE SE
-1.3904          ÉSTE Y
-0.3605          ÉSTE !EXIT
```

Figura 4.2: Bigramas obtenidos a partir del corpus DIME.

Por ejemplo, el modelo de lenguaje utilizado para el desarrollo del reconocedor de voz del proyecto DIME [Pineda, 2001] está basado en bigramas. Este modelo fue construido con el módulo *HLM* de *HTK* (ver figura 4.2). Los bigramas se construyeron a partir de la transcripción de los 31 diálogos del corpus DIME (datos de texto).

A partir de los bigramas se construye una red de palabras. Una frase válida es aquella que se puede formar al recorrer los nodos de palabras a través de los posibles arcos de la red. La *red de palabras* se construyó por medio del módulo *HBuild* de *HTK*. La figura 4.3 muestra la red de palabras generada a partir del corpus DIME, donde cada nodo representa una palabra y cada arco tiene asignada una probabilidad, esta probabilidad corresponde al bigrama formado por los dos nodos asociados a dicho arco.

Los bigramas que integran el modelo de lenguaje se construyeron a partir de los diálogos del corpus DIME. Estos diálogos se obtuvieron de grabaciones de habla espontánea en una tarea de diseño de cocinas por medio del experimento del Mago de Oz. Este experimento consistió en simular la interacción hombre-máquina por medio del habla [Villaseñor, 2000]. Por lo mismo, el sistema está limitado a reconocer sólo las palabras contenidas en el corpus DIME.

Cuando existen dos palabras acústicamente similares, el modelo de lenguaje permite seleccionar la secuencia de palabras más probable. A continuación, se presenta un ejemplo donde el sistema tiene que decidir entre las frases: "*mueve la estante*" o "*mueve la estufa*". En este ejemplo, *estante* y *estufa* son similares acústicamente. Como se ha mencionado, en un modelo de lenguaje cada palabra tiene asignada una probabilidad y esta afecta a la probabilidad de toda la se-

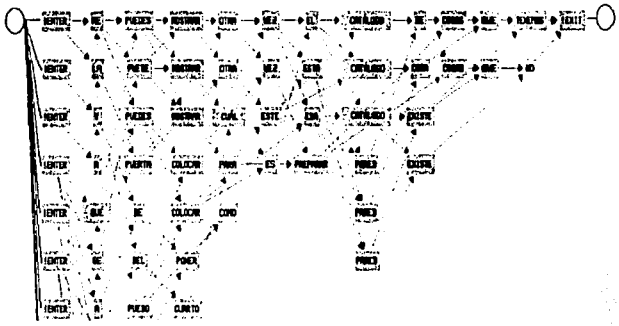


Figura 4.3: Red de palabras que representa el una parte del modelo de lenguaje.

cuencia de palabras. Si la probabilidad de *estante, P(estante).* y la probabilidad de *estufa, P(estufa)*, son iguales entonces el modelo de lenguaje se basa en la probabilidad de los bigramas *la estufa* y *la estante, P(estufa | la)* y *P(estante | la)*, respectivamente. Como es de suponerse *la estufa* es una secuencia más probable que *la estante*, es decir, *P(estufa | la) >P(estante | la)* . Por lo tanto, la probabilidad de *mueve la estufa, P(mueve la estufa)*, es mayor que la probabilidad de *mueve la estante, P(mueve la estante).* En este caso *mueve la estufa* es la secuencia de palabras más probables que obtiene el sistema de reconocimiento de voz. A continuación se formula el razonamiento descrito:

1. *P(mueve la estante)=P(mueve)P(mueve|la)P(estante|la)*
2. *P(mueve la estufa)=P(mueve)P(mueve|la)P(estufa|la)*

P(estufa|la) > P(estante|la) → P(mueve la estufa) >P(mueve la estante)

por lo tanto, las palabras reconocidas son "mueve la estufa"

En la siguiente sección se describe el diccionario de pronunciación.

4.3 Modelos de Pronunciación

Un diccionario está compuesto de un conjunto de palabras y su correspondiente pronunciación. En este trabajo de tesis, se utilizó el diccionario creado a partir

del vocabulario del corpus DIME. La pronunciación de cada palabra se representó por medio de una secuencia de símbolos fonéticos (ver tabla 4.1).

La construcción de un diccionario depende del sistema de reconocimiento de voz. Por ejemplo, si se desea construir un reconocedor específicamente para reconocimiento de dígitos, se necesita construir un diccionario que contenga un conjunto de dígitos y su forma de pronunciación, de lo contrario, si se construye un diccionario con el nombre de plantas y, lo que se requiere es reconocer dígitos, el diccionario estaría fuera de contexto. El diccionario utilizado consta de 1113 palabras diferentes, acompañadas de su transcripción fonética. El diccionario utilizado en este trabajo de tesis contiene una pronunciación por palabra. En la siguiente tabla se muestran 5 palabras del diccionario con su correspondiente pronunciación.

Palabra	Pronunciación
ABRIR	a b r i r
ABARCA	a b a r k a
ABRAN	a b r a n
ABRE	a b r e
ABREN	a b r e n

Tabla 4.1: Diccionario de pronunciación

En la siguiente sección se describen los modelos acústicos.

4.4 Modelos Acústicos

Un modelo acústico permite identificar los sonidos individuales que se producen al pronunciar una secuencia de palabras. De manera formal, el objetivo de un modelo acústico es proporcionar un método para calcular la probabilidad de observar la secuencia acústica O, cuando el hablante pronuncie la secuencia de palabras W, $P(O|W)$.

Esta probabilidad se puede estimar a partir de varias secuencias de palabras W y de la observación de sus correspondientes secuencias acústicas O. En general, esto no es práctico y en su lugar las secuencias de palabras se descomponen en las unidades acústicas correspondientes a los fonemas [Young en Uraga, 1999], donde a cada fonema le corresponde un modelo acústico.

Hay dos clases básicas de modelos acústicos sobre los cuales están basados casi todos los algoritmos de reconocimiento de voz: los modelos ocultos de Markov y las redes neuronales artificiales. Estos modelos utilizan aproximaciones

estocásticas[2] para dirigir el problema de variabilidad de la voz en el tiempo, particularmente en sistemas a gran escala. En estos modelos estocásticos se supone que la señal puede ser caracterizada como un proceso aleatorio paramétrico y que los parámetros del proceso estocástico pueden ser estimados de una manera precisa.

Los enfoques que han tenido mayor éxito en reconocimiento de voz son los que están basados en redes neuronales y modelos ocultos de Markov. En general, los modelos ocultos de Markov han demostrado tener mejor desempeño que los modelos de redes neuronales [Cole, 1995]. De los diferentes enfoques para modelar la voz, los modelos híbridos de redes neuronales y modelos ocultos de Markov han demostrado tener el mejor desempeño en reconocimiento de voz. En los casos donde la comparación ha sido controlada (que el mismo sistema ha sido usado en ambos casos) el enfoque híbrido ha logrado igual o mejor desempeño [Uraga, 1999].

A continuación; se describe el proceso para la construcción del modelo acústico basado en modelos ocultos de Markov. Este proceso consta de 3 pasos como sigue:

- Procesamiento de señales de voz
- Extracción de características acústicas
- Definición y entrenamiento de los modelos ocultos de Markov.

4.4.1 Representación y procesamiento de la señal de voz

Los datos de voz utilizados en este trabajo de tesis se grabaron a una frecuencia de muestreo de 16 KHZ (16000 muestras por segundo), donde cada muestra está representada por un número de 8 bits. Este proceso se realizó con la herramienta CSLU Toolkit. La voz producida al hablar se genera al variar la presión en el aire. Estas variaciones son continuas en el tiempo y en la magnitud, por lo que para procesar la señal de voz con una computadora es necesario realizar lo siguiente:

- Capturar la voz por medio de un micrófono y convertirla a una señal eléctrica cuya amplitud corresponda a la magnitud de la variación en la presión original.
- Muestrear la señal a determinada frecuencia, normalmente entre 8000 y 16000 Hertz. Esto permite grabar un número finito de amplitudes en un periodo de tiempo.

[2]El término aproximación estocástica se usa para indicar que los modelos que son empleados caracterizan inherentemente algo de variabilidad temporal en el habla

- Cuantizar la señal en un número discreto de bits. Con esto, cada muestra se representa con un número de bits. A esto se le llama conversión analógica-digital.

La forma de onda de una señal acústica $s(t)$ se puede representar gráficamente (ver figura 4.4). El problema en reconocimiento de voz es determinar la secuencia de palabras más probable, W dada la señal acústica $s(t)$ donde cada símbolo de W pertenece a un conjunto de unidades reconocidas $w_1(t), w_2(t), ..., w_k(t)$. Para reconocer la secuencia de símbolos dada una entrada de voz se realiza un procesamiento a la señal. El procesamiento de la señal consiste en dar una representación discreta a la señal de voz y realizar una reducción en el número de datos que contiene. Esto se realiza con el procedimiento de extracción de características [Uraga, 1999].

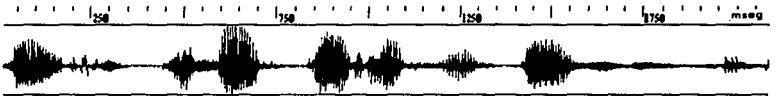


Figura 4.4: Representación gráfica de una señal de voz. El eje vertical corresponde a la amplitud de la forma de onda de la señal de voz.

4.4.2 Extracción de características

El proceso de extracción de características consiste en procesar la señal digital de voz y representarla con una secuencia de vectores de parámetros discretos. Este proceso se realiza a partir de los datos de voz; se obtiene un vector de características espectrales de un segmento o una ventana de voz con determinada duración (10 milisegundos), cada vector se integra a partir de un conjunto de datos de voz observados. Los 39 datos obtenidos fueron los siguientes:

- 12 coeficientes cesptrales en la escala de Mel (MFCC) + energía.
- la primer derivada de los 12 coeficientes cesptrales + energía (13 coeficientes delta).
- la segunda derivada de los 12 coeficientes cesptrales + energía (13 coeficientes gama).

A partir de estos datos, se obtienen 39 vectores de características con los que se puede determinar la identidad de las unidades linguísticas (en el caso de este

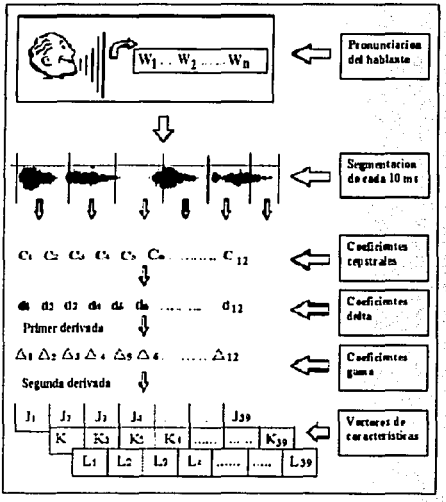


Figura 4.5: Extracción de características

trabajo, las características contenidas en cada vector determinan la identidad de cada fonema, ver figura 4.5).

Las características acústicas están muy relacionadas con la información (frecuencia, gradiente, magnitud) de los formantes de una señal de voz. Los formantes de una señal de voz se caracterizan por estar en las regiones donde la energía de la señal de voz es mayor. Las regiones de mayor energía en el espectrograma[3], que corresponden a los formantes de la señal, se transforman en las regiones más oscuras de la imagen. Normalmente se utiliza la información de los primeros tres formantes, que se nombran como F1, F2 y F3 para determinar la identidad de las unidades acústicas. En la figura 4.6 se muestra una señal de voz y sus primeros tres formantes señalados con F1, F2, F3

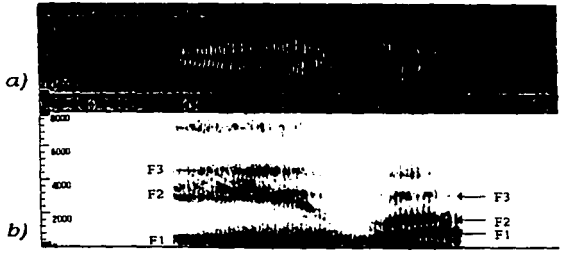


Figura 4.6: En el inciso a), se muestra la forma de onda de una señal de voz, en el inciso b) se muestran los formantes de la señal de voz (F1, F2, F3).

4.4.3 Definición de los modelos ocultos de Markov

En reconocimiento de voz cada unidad acústica se puede representar por un modelo oculto de Markov. Un modelo oculto de Markov es un conjunto de estados conectados por transiciones. Cada transición tiene dos conjuntos de probabilidades: una probabilidad de transición, que proporciona la probabilidad de tomar una transición de un estado a otro, y una función de densidad de la probabilidad de salida. Esta última es la probabilidad condicional de emitir cada

[3]Un espectrograma es una forma de representación de la señal de voz donde el eje horizontal corresponde al dominio del tiempo, el eje vertical corresponde al dominio de la frecuencia y las regiones más oscuras de la imagen corresponden a las regiones de mayor energía de la señal

símbolo de salida, a partir de un alfabeto finito, dado que se toma una transición[4] [Uraga, 1999].

En la figura 4.7 se representa gráficamente un modelo oculto de Markov comúnmente usado para representar los fonemas. Este modelo oculto de Markov está formado por tres estados etiquetados en los nodos 2, 3, 4. El estado de entrada, etiquetado en el nodo 1, y el estado de salida, etiquetado en el nodo 5, sólo son elementos que sirven para concatenar los modelos. Las ligas que tienen conectadas muestran que la probabilidad de transición del estado i al estado j es a_{ij} [Uraga, 1999].

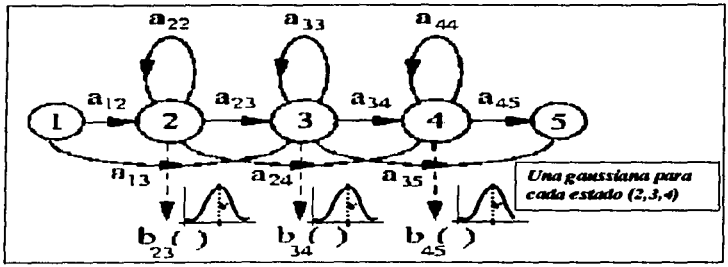


Figura 4.7: Representación de un modelo oculto de Markov.

Cada estado tiene definido un alfabeto de M símbolos que corresponden a los fonemas del español. Cada transición tiene, además, la probabilidad de emitir cada símbolo de salida dado que se toma una transición. En la figura 4.7, la probabilidad de que se genere el símbolo k dado que se tomó la transición del estado i al estado j es $b_{ij}(k)$. Un modelo oculto de Markov se define por los siguientes parámetros:

- N es el número de estados en el modelo. Los estados individuales se denotan como $S = S_1, S_2, ..., S_N$ y el estado en el tiempo t como X_t.

  En reconocimiento de voz cada estado representa las unidades acústicas de reconocimiento. Generalmente, un fonema se divide en 3 partes: las partes inicial, central y final.

- M es el número de símbolos distintos de observación por estado, i.e., el tamaño del alfabeto. Los símbolos individuales se denotan como $V = v_1, v_2, ..., v_M$. El símbolo de observación en el tiempo t se denota como o_t y

[4] En la mayoría de la literatura, la función de densidad de la probabilidad de salida, bij, se asocia con los estados y no con las transiciones entre estados.

corresponde a uno de los elementos de V. En reconocimiento de voz estos símbolos de observación corresponden a los fonemas o símbolos fonéticos utilizados en la transcripción fonética de los datos de voz.

- La distribución de probabilidad de transición, $A = a_{ij}$, donde a_{ij} es la probabilidad de tomar una transición del estado i al estado j. a_{ij} está relacionada con la probabilidad de observar una unidad acústica correspondiente a S_j dado que otra unidad acústica correspondiente a S_i fue observada.

- La distribución de probabilidad de emitir un símbolo v_k cuando se toma la transición del estado i al estado j, $B = b_{ij}(v_k)$. $b_{ij}(v_k)$ está relacionada con la probabilidad de que al cambiar a S_j se observe el símbolo correspondiente a v_k.

  La función $b_{ij}(v_k)$ se aproxima con distribuciones gaussianas. Una función de distribución gaussiana se puede definir a partir de dos parámetros: la media y la varianza de los datos de voz observados.

4.4.4 Entrenamiento de Modelos Ocultos de Markov

Para entrenar los modelos acústicos se utilizó un corpus de voz. Este *corpus* contiene 1515 frases grabadas por 50 hablantes. Una oración consta aproximadamente de 4 segundos de duración, por lo tanto, se tienen aproximadamente 2 horas de grabación. Las frases grabadas contienen toda la variedad fonética para el español hablado en México; además, se procuró obtener una distribución fonética en proporciones aproximadas a la frecuencia de aparición en el habla.

El proceso de entrenamiento se ilustra gráficamente en la figura 4.8 Este proceso de entrenamiento se realiza con el algoritmo *forward-backward* (también conocido como *Baum-Welch*)[Young, 1997]. Este algoritmo está implementado en el conjunto de programas de la herramienta (HTK) que fue utilizado en este trabajo de tesis para crear los modelos acústicos basados en modelos ocultos de Markov para cada fonema.

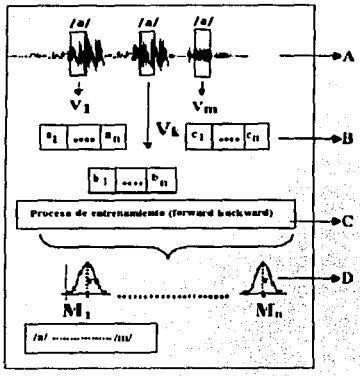


Figura 4.8: Proceso de entrenamiento de Modelos acústicos. A) Datos de voz, B) Vectores de características extraídos de los datos de voz, C) Proceso de entrenamiento de modelos acústicos por medio del algoritmo forward-backward, D) Conjunto de medias y varianzas obtenidas a partir de los vectores de características del corpus de voz. Se obtuvieron 39 gaussianas para cada fonema a partir de los datos de voz.

4.4.5 Reconocimiento

El proceso de reconocimiento de voz se realiza mediante un proceso de búsqueda que utiliza el algoritmo de *Viterbi*. Este algoritmo está implementado en el programa *HVite* de la herramienta HTK [Young, 1997]. Este proceso se representa gráficamente en la figura 4.9.

4.5 Integración de los modelos ocultos de Markov en el SRV

El nuevo modelo acústico se integró al sistema de reconocimiento de voz del sistema creado en el proyecto DIME. La figura 4.10 ilustra la arquitectura del sistema de reconocimiento de voz. El reconocedor es independiente del hablante, y por lo mismo, se puede utilizar para cualquier aplicación que utilice reconocimiento de voz continua; es decir, se presume que puede obtenerse un alto porcentaje de reconocimiento en una gran cantidad de hablantes. Finalmente, el sistema se

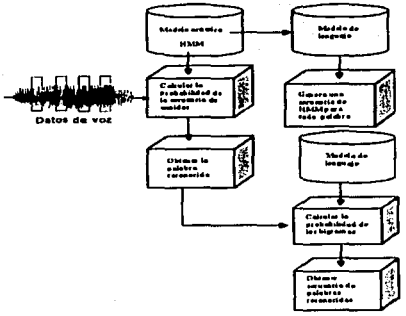


Figura 4.9: Proceso de reconocimiento de voz.

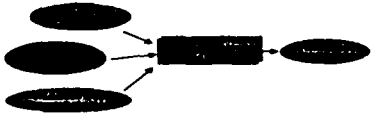


Figura 4.10: Arquitectura del nuevo sistema de reconocimiento de voz

evaluó por medio de los programas *HVite* y *HResults* de la herramienta HTK. Este proceso se describe en el siguiente capítulo.

4.6 Conclusiones

En este capítulo se planteó formalmente el problema de reconocimiento de voz. Se describió cada uno de los módulos que integran un reconocedor poniendo énfasis en los modelos acústicos. De manera muy general, se describieron los procesos de entrenamiento y reconocimiento de voz. Finalmente, se describió la arquitectura del nuevo sistema de reconocimiento de voz.

Capítulo 5

Pruebas y Resultados

En este capítulo se presentan los resultados obtenidos al evaluar el sistema de reconocimiento de voz con tres modelos acústicos distintos. Se presentan la forma de evaluar el reconocedor de voz y el desempeño alcanzado por el sistema.

5.1 Método de evaluación del reconocedor

Una vez que el sistema de reconocimiento de voz ha obtenido las secuencias de palabras reconocidas, el siguiente paso es analizar los resultados. La herramienta *HResults* de HTK es la encargada de cumplir dicha misión; compara las secuencias de palabras reconocidas contra las transcripciones de referencia y finalmente, obtiene el número de errores por substitución (S), omisión (O) e inserción (I). El porcentaje de reconocimiento se calcula mediante la siguiente fórmula:

$$EP = 100 - \%O - \%S - \%I \text{[Young, 1997]} \tag{5.1}$$

donde EP es el porcentaje exacto de palabras reconocidas.

5.2 Pruebas

En esta sección se presentan las pruebas realizadas con los tres modelos acústicos integrados en el mismo sistema de reconocimiento de voz. Los modelos acústicos creados se evaluaron utilizando las frases de los diálogos 1, 5, 6 y 12 del corpus DIME.

La primera prueba se realizó con el modelo acústico M1. Este modelo fue creado a partir de una parte del corpus de voz grabada por 20 de los 50 hablantes. El desempeño del sistema de reconocimiento de voz utilizando el modelo M1 se

muestra en la siguiente tabla:

Frases	Palabras	O	S	I	%Palabras Reconocidas
612	2657	382	654	34	59.73%

Tabla 5.1: Evaluación del sistema utilizando el modelo M1.

La segunda prueba se realizó con el modelo acústico M2. Este modelo fue creado a partir de una parte del corpus de voz grabada por 40 de los 50 hablantes. El desempeño del sistema de reconocimiento de voz utilizando el modelo M2 se muestra en la siguiente tabla:

Frases	Palabras	O	S	I	%Palabras Reconocidas
612	2657	446	721	32	54.87%

Tabla 5.2: Evaluación del sistema utilizando el modelo M2.

La tercera prueba se realizó con el modelo acústico M3. Este modelo fue creado a partir del corpus de voz completo grabado por 50 hablantes. El desempeño del sistema de reconocimiento de voz utilizando el modelo M3 se muestra en la siguiente tabla:

Frases	Palabras	O	S	I	%Palabras Reconocidas
612	2657	287	478	24	70.30%

Tabla 5.3: Evaluación del sistema utilizando el modelo M3.

5.3 Conclusiones

En este capítulo se presentaron los resultados obtenidos durante la evaluación del sistema de reconocimiento de voz. Este sistema fue evaluado utilizando tres modelos acústicos diferentes. El máximo resultado alcanzado fue de 70.23% de reconocimiento continuo de palabras. Este resultado muestra que con los nuevos modelos acústicos se logró un incremento considerable en el desempeño del sistema de reconocimiento de voz.

Capítulo 6

Conclusiones y trabajo a futuro

El objetivo general de este trabajo de tesis fue modelar acústicamente la voz del español hablado en México, utilizando un enfoque basado en modelos ocultos de Markov, con el propósito de construir un sistema de reconocimiento de voz continua e independiente del hablante.

Para cumplir dicho objetivo, fue necesario diseñar y obtener un *corpus* de voz rico y balanceado fonéticamente. Con este *corpus* de voz, se construyó un nuevo conjunto de modelos acústicos para reconocimiento de voz continua e independiente del hablante para el Español hablado en México. Estos nuevo modelos acústicos se integraron al módulo de reconocimiento de voz del proyecto DIME. El objetivo de mejorar el desempeño del sistema de reconocimiento de voz se cumplió satisfactoriamente al incrementar en un 20% el desempeño del sistema anterior.

En este trabajo de tesis se describió qué es un sistema de reconocimiento de voz, cuál es su arquitectura y los modelos principales que lo integran: un modelo acústico, un modelo de lenguaje y un modelo de pronunciación. También se describió qué es un *corpus* y se propuso una metodología para crear un *corpus* de voz rico y balanceado fonéticamente. Este corpus sirvió como base para la creación de un conjunto de modelos acústicos para reconocimiento de voz continua e independiente del hablante. Estos modelos acústicos se crearon mediante una herramienta computacional (HTK) que se basa en el enfoque de modelos ocultos de Markov (HMM). Finalmente; se presentaron los resultados de las pruebas realizadas al sistema de reconocimiento de voz con los nuevos modelos acústicos.

Las aportaciones principales de este trabajo de tesis fueron las siguientes:

- *Un corpus de voz rico y balanceado fonéticamente*: Este *corpus* fue grabado con 50 hablantes mexicanos, cuyos datos de voz contienen a todos los fonemas del español en una proporción aproximada a su frecuencia de aparición en el habla.

- *Un nuevo conjunto de modelos acústicos*: El modelo acústico creado consta de un modelo oculto de Markov para cada fonema del español. Este modelo acústico permite identificar, de manera general, los sonidos del español hablado en México.Estos modelos se crearon a partir del *corpus* rico y balanceado. Estos modelos acústicos pueden ser utilizados para desarrollar diferentes sistemas que utilicen reconocimiento de voz en español independientemente de la aplicación.

- *Un sistema de reconocimiento de voz continua e independiente del hablante*: Este sistema reconoce la voz de un número ilimitado de hablantes. Su aplicación es de dominio general; es decir, puede utilizarse para cualquier tipo de aplicación. Cabe resaltar que este sistema se está utilizando actualmente en el proyecto " Navegación de un robot móvil por medio de visión computacional y de lenguaje natural"[1], su desempeño actual ha tenido resultados muy satisfactorios.

6.1 Objetivos alcanzados

Los objetivos de este trabajo de tesis se lograron satisfactoriamente y se presentan a continuación:

- *Plantear y seguir una metodología para crear un corpus de voz.*
- *Crear modelos acústicos-fonéticos independientes del hablante.*
- *Integrar los modelos acústicos en el sistema de reconocimiento de voz.*
- *Evaluar el sistema de reconocimiento de voz.*

A continuación se describe como se lograron estos objetivos :

6.1.1 Metodología para crear un corpus de voz

Un *corpus* de voz es un recurso indispensable para desarrollar tecnología de lenguaje hablado (reconocimiento o síntesis de voz) sin embargo; existen muy pocos recursos linguísticos disponibles, por lo mismo; en este trabajo de tesis se planteó como uno de sus principales objetivos, desarrollar una metodología para crear un *corpus* de voz.

La metodología consitió en diseñar el contenido linguístico del *corpus*, recolectar texto,

[1]Este proyecto también está siendo desarrollado por el grupo de inteligencia artificial del IIMAS.

transcribir de texto a fonemas, seleccionar frases, grabar las frases seleccionadas y transcribir los datos de voz. Este proceso permitió obtener un *corpus* de voz con las siguientes características:

- *Datos de voz continua en español.* Se obtuvieron a partir de 1515 frases leídas por 50 hablantes (25 hombres y 25 mujeres).
- *Datos de voz transcritos a nivel ortográfico y fonético.*
- *Riqueza fonética.* El *corpus* de voz, en su variedad linguística, contiene todos y cada uno de los fonemas del Español.
- *Balanceo fonético.* El *corpus* de voz contiene una variedad fonética aproximada a la frecuencia de aparición de cada fonema en el habla. Este *corpus* fue utilizado para crear modelos acústicos para un sistema de reconocimiento de voz. Las dimensiones del *corpus* se presentan en la tabla 6.1.

Tiempo de Grabación	2 horas
Número de Palabras	14204
Número de Fonemas	86010

Tabla 6.1: Características del corpus.

Comparando la duración de algunos corpus (de 30 a 200 horas), es muy claro notar que a pesar de haber generado un corpus con tan pocos datos (2 horas de grabación) el desempeño alcanzado por nuestro sistema es similar al de los sistemas creados con grandes cantidades de datos. Esta ventaja es un consecuencia directa de haber diseñado adecuadamente el contenido linguístico del corpus.

6.1.2 Creación de modelos acústicos-fonéticos

En este trabajo de tesis, se crearon modelos acústicos basados en modelos ocultos de Markov a partir de un corpus rico y balanceado fonéticamente. Se obtuvo un modelo oculto de Markov por cada fonema del Español. Cada modelo acústico se entrenó a partir de los datos de voz del *corpus*; esta tarea se realizó mediante la herramienta Hidden Markov model Tool Kit (HTK). Estos modelos se integraron en un sistema de reconocimiento de voz.

6.1.3 Sistema de Reconocimiento de Voz

El sistema de reconocimiento de voz creado en este trabajo de tesis, fue integrado al sistema del proyecto DIME. Con este trabajo se mejoró el desempeño del

módulo de reconocimiento de voz. El nuevo sistema de reconocimiento de voz se formó con los nuevos modelos acústicos y con el modelo de lenguaje y el modelo de pronunciación del sistema anterior.

6.1.4 Evaluación del Sistema

Los resultados obtenidos demuestran que los modelos acústicos creados a partir de un corpus de voz rico y balanceado fonéticamente mejoraron considerablemente el desempeño del sistema de reconocimiento de voz. Los resultados que se habían obtenido con el sistema anterior eran de 50% para reconocimiento de voz continua. Con la integración de los nuevos modelos acústicos se incrementó el desempeño del sistema en un 20%; es decir, se alcanzó un 70.23% de reconocimiento de voz continua.

6.2 Trabajo a futuro

El desempeño del sistema de reconocimiento de voz se puede mejorar de la siguiente manera:

- Incluyendo una mayor cantidad de datos de voz en el *corpus* para crear nuevos modelos acústicos que permitan crear un sistema más robusto ante un mayor número de hablantes, ruido ambiental, etc.
- Ampliando el vocabulario del corpus y modelando su pronunciación con información alofónica que represente las distintas formas de pronunciar una palabra.
- Incorporando modelos de ruidoal sistema de reconocimiento de voz. Los errores causados por ruido excesivo pueden ser disminuidos incorporando nuevos modelos de ruido ambiental y de ruido producido por el hablante (toser, reír, estornudar, respirar,etc.).

Bibliografía

[Ahuactzin, 1999] Ahuactzin (1999), *Reconocimiento y Sintesis de Voz*

[Beulen, 1998] Beulen (1998), *Pronunciation modelling in the RWTH large vocabulary speech recognizer.*

[Canfield,1992] Canfield, (1992), *Spanish pronunciation in the Americas*, The University of Chicago Press, U.S.A.

[Casacuberta, 1987] Casacuberta (1987), *Reconocimiento Automático del Habla*, Marcombo, Barcelona-México.

[Charniak, 1993 en Uraga, 1999] Uraga (1999), *Modelado fonético para un Sistema de Reconocimiento de Voz continua en Español*

[Cole, 1995] Cole, R., Hirschman, L., Atlas, L., Beckman, M., Biermann, A., Bush, M., Clement, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N. Novick, D. G., Ostendorf, M., Oviatt, S., Price, P., Silverman, H. Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., Zue, V., (1995), *The Challenge of Spoken Language Systems Research Directions for the Nineties, IEEE Transactions on Speed and Audio Processing*, Vol.3, No.1, enero de 1995.

[Cole en Uraga, 1999] Uraga (1999) *Modelado fonético para un Sistema de Reconocimiento de Voz continua en Español*

[CSLU, 2002] CSLU (2002) *The CSLU Toolkit.* URL:cslu.cse.ogi.edu/cslu/

[CSTAR II, 2002] CSTAR II (2002) *Project CSTAR II.* URL:www.speech.cs.cmu.edu/CSTARII/

[Cosi, 1998] Cosi (1998), *CSLU*, URL:cslu.ogi.cse.edu/cslu/documentation

[Crystal, 1991] Crystal (1991), *A Dictionary of Linguistics and Phonetics, Blackwell*, 3rd Edition.

[Dialogic, 2002] Dialogic, URL:http://www.dccusa.com DCC Solutions for security

[Delattre, 1965] Delattre (1965), *Comparing the phonetic features of English, German, Spanish and French*, Julius Groos Verlag, Santa Barbara California.

[Dragon Systems, 2002] Dragon Co. (2002), *Naturally Speaking*, URL:www.scansoft.com/naturallyspeaking/

[Garcia, 1994] García M., J., (1994), *Reconocimiento de voz usando redes neuronales.*, Tesis de maestría, ITESM Campus Morelos, Cuernavaca, Morelos, México.

[Gibbon en Uraga, 1999] Uraga (1999), *Modelado fonético para un Sistema de Reconocimiento de voz continua en Español*

[Jelinek, 1985, Young, 1996 en Uraga, 1999] Uraga (1999) *Modelado fonético para un Sistema de Reconocimiento de Voz continua en Español*

[Kirschning, 2001] Kirschning (2001), *Procesamiento Automático de Voz*, URL: http://mailweb.udlap.mx/ ingrid/

[Kirschning, 2000] , *Aplicación de tecnología de voz en la enseñanza del español*, in Proc. of the 1er. Taller Internacional de Tratamiento del habla, procesamiento de voz y el lenguaje, HAVOL 2000, Mexico, Julio 2000.

[Listerri, 1999] Listerri (1999), *Filología e Informática*, Barcelona.

[Microsoft-MST, 2002] Microsoft (2002), URL:http://research.microsoft.com/srg/ Microsoft Speech Tecnology

[Morgan, 1991 en Fragoso, 2001] Fragoso (2001), *Entrenamiento y comparación de un nuevo reconocedor basado en redes neuronales y modelos ocultos de Markov*, Tesis de Licenciatura, Universidad de las

Americas, Puebla.

[Munguía, 1998] Munguía (1998), *Gramática, Lengua Española, Reglas y Ejercicios*, Larousse, México.

[Munive,1999] Munive,Vargas,Serridge, Cervantes, Kirschning, *Un sistema de correo electrónico y de voz usando reconocimiento de voz*, Soluciones avanzadas, 7, 69, (Mayo 1999), 44-48.

[Munive, 1999] *Entrenamiento de un reconocedor fonético de dígitos para el español de México usando el CSLU Toolkit*,Computación y sistemas, 3,2, (1999), 98-104.

[Narada Warakagoda,1996] Warakagoda (1996), *Hidden Markov Models*, URL:http:jedlik.phy.bme.hugerjanosHMMnode2.html.

[Novell, 2002] Novell (2002), *Technical information report*, URL:http://support.novell.com/servlet/tidfinder/1203747 Novell Technology

[Olivier, 1999] Olivier (1999), *Evaluación de métodos de determinación automática de una transcripción fonética*, Tesis de Licenciatura, Universidad de las Americas, Puebla.

[Pineda, 2001] Pineda, (2001), *Dime A multimodal project for kitchen design*

[Polyglot, 2002] Proyecto Polyglot, (2001), URL:www.newcastle.research.ec.org./esp-syn/text/2104.html

[Radová, 1999] Radová (1999), *Methods of Sentences Selection for Read-Speech Corpus Design*

[RAE, 1992] Real Academia de la lengua Española (1992), Diccionario Electrónico, URL:http:www.rae.es

[Rojo, 1991] Rojo (1991), *Frecuencia de fonema en español actual*, in BREA, Santiago de Compostela.

[Seller en Uraga, 1999] Uraga (1999), *Modelado fonético para un Sistema de Reconocimiento de voz continua en Español*

[Sensory Circuits, 2002] Sensory Circuits, URL:http://www.sensoryinc.com

[Sinclair en Listerri, 1999] Listerri (1999), *Filología e Informática*

[Strik, 1998] Strik (1998), *Modeling Pronunciation variation for ASR: Overview and comparison of methods*

[Sunburst, 2002] Sunburst, URL:http://www.sunburstmedia.com/CSW.html

[Timit, 1993] Garofolo John, Lori F.Lamel, William M.Fisher, (1993), URL:http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[Tlatoa, 1998] Tlatoa (1998), *Speeh Technology Research Group: TLATOA, URL: info.pue.udlap.mx/ sistemas/tlatoa/*

[Uraga, 1999] Uraga (1999), *Modelado fonético para un Sistema de Reconocimiento de Voz continua en Español*, Tesis de Maestría, Instituto Tecnológico de Monterrey, Campus Morelos.

[Uraga y Pineda, 2002] Uraga, Pineda (2002), *Automatic Generation of Pronunciation Lexicons for Spanish*, Third Internation Conference, CICLing 2001, Mexico.

[Vaufreydaz, 1999] Vaufreydaz, Akbar, Rouillard (1999), *Internet Documents: A rich source for Spoken Languague modeling*

[Villaseñor, 2000] Villaseñor (2000), *The Dime Corpus*, IIMAS-UNAM, Lecture Notes in Computer Science.

[Vocalis Co.,2002] Vocalis, URL:www.vocalis.com Voice driven business solutions.

[Wang, 1999] Wang, Shing (1999), *An algorithm for Automatic generation of mandarin phonetic balanced corpus*

[Wooters, 1996] Wooters (1996), *Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System*, Congress.

[Yannakoudakis, 1985] Yannakoudakis, Hutton (1985), *Speech synthesis and Recognition Systems*, Elis Horwook Limited, England.

[Young en Uraga, 1999] Young (1996), *Large vocabulary Continous Speech Recognition*: a review Report, Cambridge University Engineering Department.

[Young, 1997] Young (1997), *The HTK Book*, Cambridge University, 1997.

Apéndice A

Programa que extrae texto de internet

```
#!/usr/bin/perl

if (@ARGV  != 1) {
    print "ERROR: wrong number of arguments\n";
}

($file)=@ARGV;

@m= ("01","02","03","04","05","06","07","08","09","10","11","12");

@d= ("01","02","03","04","05","06","07","08","09","10","11","12",\\
"13","14","15","16","17","18","19","20", "21","22","23","24","25",\\
"26","27","28");

open(FECHAS,">$file.fec") || die "NO puedo crear $file.fec\n";

#nov-1999 a dic-1999
$Nmes=10;
foreach $mes (nov,dic) {
    for ($dia=0; $dia<28; $dia++) {
print FECHAS "http://www.jornada.unam.mx/1999/$mes";
print FECHAS "99/99$m[$Nmes]$d[$dia]/\n";
    }
    $Nmes++;
}
#ene-2000 a dic-2000
$Nmes=0;
foreach $mes ("ene","feb","mar","abr","may","jun","jul","ago",\\
```

```
"sep","oct","nov","dic") {
    for ($dia=0; $dia<28; $dia++) {
print FECHAS "http://www.jornada.unam.mx/2000/$mes";
print FECHAS "00/00$m[$Nmes]$d[$dia]/\n";
    }
    $Nmes++;
}
#ene-2001 a dic 2001
$Nmes=0;
foreach $mes (ene,feb,mar,abr,may,jun,jul,ago) {
   for ($dia=0; $dia<28; $dia++) {
       print FECHAS "http://www.jornada.unam.mx/2001/$mes";
       print FECHAS "01/01$m[$Nmes]$d[$dia]/\n";
    }
   $Nmes++;
}
close(FECHAS);

# "-i file" Extrae los URL desde un archivo "file"
# "-r -l 2" Extrae URL's recursivamente hasta 2 niveles
# "-L"      Extrae unicamente lo que este en el mismo servidor
# "-A html,txt" Extrae unicamente html y txt
# "-Q 200M"     Extrae un maximo de 200 Megas en arhivos
# "-nc"         Extrae solo los archivos que faltaron en un
# "-nH"  Crea el directorio con el nombre de la pagina http \\
y ahi guarda los archivos de la busqueda
# "-m"   Para crear un espejo del sitio

system("wget -i $file.fec -m -r -L -A html,txt,htm -Q 200M -nH");

#Para generar un mirror:
#wget -m -r -L -A html,txt -Q 200M -nH http://www.jornada.unam.mx
```

Apéndice B

Programa que convierte de formato HTML a TXT

```
#!/usr/bin/perl

#Improved HTML2TXT
#Sintaxis: html2txt listahtml

while(<>) {
  #$fn = $ARGV[0] unless ! $ARGV[0];

  #if (!$ARGV[0]){
  #    print "Input File:\n";
  #    chop($fn = <STDIN>);
  #    }

  $fn=$_;
  open (INF,"< $fn");
  $fn=~ s/\.htm?./\.txt/;
  print "$fn\n";
  open (OUF,"> $fn");

  sub par
  {
      $par = shift;
      $par =~ s/\<p\>/\./g;
      $par =~ s/<.*script.*>//gsi;
      $par =~ s/<img.+>//gsi;
      $par =~ s/<br>/\./gs;
      $par =~ s/\<\/.*?\>/\./gs;
```

```
$par =~ s/<.*?>//gs;

$par =~ s/\&a(grave;|acute;)//g;
$par =~ s/\&e(grave;|acute;)//g;
$par =~ s/\&i(grave;|acute;)//g;
$par =~ s/\&o(grave;|acute;)//g;
$par =~ s/\&u(grave;|acute;)//g;
$par =~ s/\&A(grave;|acute;)//g;
$par =~ s/\&E(grave;|acute;)//g;
$par =~ s/\&I(grave;|acute;)//g;
$par =~ s/\&O(grave;|acute;)//g;
$par =~ s/\&U(grave;|acute;)//g;
$par =~ s/\&ntilde;//g;
$par =~ s/\&Ntilde;//g;
$par =~ s/\&uuml;//g;
$par =~ s/\&Uuml;//g;
$par =~ s/(\&)(\w)(grave;|acute;)/$2\'/g;

$par =~ s/\ /\./g;
$par =~ s/\&lt;/</g;
$par =~ s/\&gt;/>/g;
$par =~ s/\&quot;/\"/g;
$par =~ s/\&#168;/\"/g;

$par =~ s/\&#173;/\-/g;
$par =~ s/\&shy;/\-/g;

$par =~ s/\&#191;/\/g;
$par =~ s/\&iquest;/\/g;

$par =~ s/\&#161;/\/g;
$par =~ s/\&#1excl;/\/g;

$par =~ s/\&#164;/\*/g;
$par =~ s/\&curren;/\*/g;

$par =~ s/\&#186;/\/g;
$par =~ s/\&ordm;/\/g;
$par =~ s/\&#170;/\/g;
$par =~ s/\&#176;/\/g; #es simbolo de grados

$par =~ s/&#180;/'/g;

$par =~ s/\&#130;/\,/g;
```

```
    $par =~ s/\&#009;//g;
    $par =~ s/\&#131;//g;
    $par =~ s/\&#149;//g;

    $par =~ s/\&#174;//g;  #simbolo de marca registrada

    $par =~ s/\&#[0-9]+;//g;

    $par =~ s/^\s*$//g;

    # lineas que modifique del programa mostrado
    #      $par =~ s/<.*script.*>;
    #      $par =~ s/<img.+>/\n---------\n\| \\
    Image \|\n---------\n/gsi;
    #      $par =~ s/<br>/\n/gs;

 }

 while ($nl=<INF>)
 {
 $cl .= $nl;
    if ($cl =~ /.*>[^<]*\n/)
    {
    par($cl);
    print OUF $par;
    undef $cl;
    }
 }

 par($cl);
 print OUF $par;
 close (INF);
 close (OUF);
}
```

Apéndice C

Programa que convierte cifras numricas a palabras

```
#!/usr/bin/perl

#Este programa procesa un conjunto de textos y selecciona


#sintaxis: cuentafrases lista-archivos frases-selec
otras-frases acronimos dicc
#entrada: lista-archivos
#salida: frases-selec otras-frases acronimos


if (@ARGV  != 5) {
    print "ERROR: wrong number of arguments\n";
}
($file_list,$seleccionadas, $otras, $acro, $dicc)=@ARGV;

# Archivo que contiene la lista de textos
que seran procesados:
open (LIST, $file_list) ||  die
 "No puedo abrir el archivo $file_listdir";

# Archivos de salida
open (FRASES, ">$seleccionadas") ||
die "No puedo abrir el archivo $seleccionadas";
open (OTRAS,">$otras") || die
"No puedo abrir el archivo $otras";
open (ACRO,">$acro") || die
"No puedo abrir el archivo $acro";
```

```
leeDiccionario($dicc);

while (<LIST>) {
    s/\n//g;
    $file_txt=$_;
    open (FILE, $file_txt) || die
"No puedo abrir el archivo $file_txt";
    $linea="";

    # Concatena las lineas y parrafos de un archivo
    while (<FILE>) {
s/\n//g;
if ($linea=!~\|\AAPAUNAM|\ AARC|\ AAT|\
 AATF|\ AB|\ ABA1|\ ABAC|\ ABB|\ ABBA|\ ABB|\
ABC|\ ABCDE|\ ABEDROP|\ ABFA|\ ABIA|\ ABM|\
ABMAFP|\ ABMREUTERS|\ ABN|\ ABNAMOR|\
ABNAMRO|\ AC|\ ACA|\ ACAAN|\ ACAS|\
ACAT|\ ACBN|\ ACC|\ ACCIONESS|\
ACCU|\ ACDB|\ ACDC|\ ACDDH|\
ACDH|\ ACE|\ ACEM|\ ACEX|\ ACF|\
ACFV|\ ACGA|\ ACH|\ ACHACH|\
ACI|\ ACIR|\ ACJM|\ ACLAN|\
ACLU|\ ACM|\ ACME|\ ACMM|\
ACN|\ ACNO|\ ACNR|\ ACNUR|\
ACO|\ ACOPECH|\ ACP|\ ACPC|\
ACPE|\ ACPT|\ ACROBAT|\ ACT|\
ACTUP|\ AD|\ ADAI|\ ADAN|\
ADC|\ ADCEBRA|\ ADDY|\
ADEGI|\ ADHOC|\ ADI|\ ADIAT|\
ADIIFH|\ ADIVAC|\ ADM|\ ADN|\
ADNM|\ ADNRNA|\ ADNY|\ ADO|\
ADP|\ ADR|\ ADS|\ AE|\ AECI|\
AEDS|\ AEEU|\ AEGIS|\ AEI|\
AEK|\ AELC|\ AEMBDF|\ AENA|\
AEPCP|\ AEPF|\ AES|\ AF|\ AFA|\
AFADEM|\ AFANES|\ AFC|\ AFDD|\
AFEP|\ AFF|\ AFFAIRS|\ AFI|\
AFIP|\ AFIS|\ AFISA|\ AFLCCIO|\
AFLCIO|\ AFM|\ AFP|\ AFPCOPA|\
AFPDEFINEN|\ AFPJUSTICIA|\
AFPTEIXEIRA|\ AFPZIDANE|\
AFSCC|\ AFSCME|\ AFT|\ AFTRA|\
AG|\ AGA|\ AGEB|\ AGEU|\ AGF|\
AGI|\ AGN|\ AGOPSTO|\ AGP|\
```

```
AGT|\ AHISA|\ AHISTRICO|\
AHMDF|\ AHMSA|\ AHOLD|\
AHP|\ AHPS|\ AHPSP|\ AHQ|\
AI|\ AIA|\ AIB|\ AIBA|\ AIBO|\
AICM|\ AID1|\ AIDOH|\ AIDS|\
AIE9|\ AIEM|\ AIFI|\ AIFI|\
AIHR|\ AIM|\ AIMUR|\ AIN|\
AIP|\ AIPAC|\ AIPM|\ AISA|\
AISF|\ AIU|\ AIUSA|\ AIX|\
AJ|\ AJA|/)
$linea= "$linea $_";
    }

    $linea= sustituyeAcronimos($linea);
    $linea= sustituyeNumeros($linea);

    # Separa las lineas de texto por punto
    (@parrafos)= split(/\.|\,|\;/, $linea);
    foreach $_ (@parrafos) {
$num=0;
s/\(|\)|\$|\:|\/|\"|\'|\'|\#|\*|\-|\.|\?
|\|\|\!|\[|\]|\&|/ /g;
s/^\s+//g; s/\s+/ /g;  #Eliminar espacios sobrantes

if ($_=~/[a-zA-ZO-9]+/)
{
    (@lista)= split(/\s+/,$_); #Cuenta las palabras
    foreach $palabra (@lista) {
$num++;
# Obtiene acrnimos
if ( ($palabra=~/[A-Z]+/) &&
(!($palabra=~/[a-zO-9]+/))) {
    $acronimo{$palabra}++;
    $numacro++;
}
    }
    $total++;

    $numpalabras{$num}++;

    #SI LA FRASE TIENE ENTRE 5 Y 15 PALABRAS, SE GUARDA
    if (($num>4) && ($num<25)){
print FRASES "$_\n";
$selec++;
    } else
```

```
    {
print OTRAS "$num.$_\n";
$noselec++;

#if ($num>100) {
print "Mas de 50: $file_txt\n";}
    }
}
close (FILE);
    }
}
close(DIRS);
close(FRASES);
close(OTRAS);

foreach $palabra (sort keys(%acronimo)) {
print ACRO "$palabra\t$acronimo{$palabra}\n";
    }
close(ACRO);

print "$selec de $total frases fueron seleccionadas
($noselec frases no fueron seleccionadas).\n";
print "$numacro acrnimos en el texto\n";

foreach $numero (sort bynum keys(%numpalabras)) {
print  "Frases con $numero palabras: $numpalabras{$numero}\n";
}

#FUNCION QUE ORDENA ELEMENTOS
sub bynum { $a <=> $b; }


#Diccionario que contiene el nombre de ciertas
palabras abreviadas Ej. D.F. Distrito Federal
sub leeDiccionario {
    ($diccFile)=@_;
    local($simb,@palabras);
    open (DICC, $diccFile) ||  die
 "No puedo abrir el archivo $dicc";
    while(<DICC>)
    {
($simb,@palabras)= split(/\t/, $_);
$diccionario{$simb}= join(" ",@palabras);
    }
    close(DICC);
```

```
}

#Sustituye el nombre de los acrnimos en una frase
sub sustituyeAcronimos {
    ($_)=@_;
    local($simb);

    foreach $simb (keys(%diccionario)) {
s/$simb/$diccionario{$simb}/g;
    }
    return($_);
}

sub sustituyeNumeros {
    ($_)=@_;
    local(@lista,$palabra,$letras);

    (@lista)= split(/\s+/,$_); #obtiene las palabras
    foreach $palabra (@lista) {

if ($palabra=~/[0-9]+/){
    $palabra=~s/\(|\)|\$|\:|\/|\"|\'|\'|\#|\*
|\-|\,|\;|\.|\?|\|\|\|\[|\]/ /g;
    $letras=transcribeNumeros($palabra);
    s/$palabra/$letras/;
}

    }
    return($_);
}

sub transcribeNumeros {
    local($numero)=@_; #posicion de derecha a izquierda
    local($letras, @digitos,$dig,$decenas,$long);

    $letras="";
    $long= length($numero);
    @digitos=split(//,$numero);

    foreach $dig (@digitos)
    {


    if ($long==4) {
if ($dig==1) {$letras= "$letras mil"; }
```

```
if ($dig==2) {$letras= "$letras dos mil"; }
if ($dig==3) {$letras= "$letras tres mil"; }
if ($dig==4) {$letras= "$letras cuatro mil"; }
if ($dig==5) {$letras= "$letras cinco mil"; }
if ($dig==6) {$letras= "$letras seis mil"; }
if ($dig==7) {$letras= "$letras siete mil"; }
if ($dig==8) {$letras= "$letras ocho mil"; }
if ($dig==9) {$letras= "$letras nueve mil"; }
    }

    if ($long==3) {

if ($dig==1) {$letras= "$letras ciento" };
if ($dig==2) {$letras= "$letras doscientos" };
if ($dig==3) {$letras= "$letras trescientos" };
if ($dig==4) {$letras= "$letras cuatrocientos" };
if ($dig==5) {$letras= "$letras quinientos" };
if ($dig==6) {$letras= "$letras seiscientos" };
if ($dig==7) {$letras= "$letras setecientos" };
if ($dig==8) {$letras= "$letras ochocientos" };
if ($dig==9) {$letras= "$letras novecientos" };
    }

    if ($long==2) {
$decenas=$dig;
if ($dig==3) {$letras= "$letras treinta" };
if ($dig==4) {$letras= "$letras cuarenta" };
if ($dig==5) {$letras= "$letras cincuenta" };
if ($dig==6) {$letras= "$letras sesenta" };
if ($dig==7) {$letras= "$letras setenta" };
if ($dig==8) {$letras= "$letras ochenta" };
if ($dig==9) {$letras= "$letras noventa" };

    }

    if ($long==1) {
        if (($decenas>=3) && ($dig!=0)) { $letras= "$letras y"; }

if ($decenas==1) {
    if ($dig==0) {$letras= "$letras diez" };
    if ($dig==1) {$letras= "$letras once" };
    if ($dig==2) {$letras= "$letras doce" };
    if ($dig==3) {$letras= "$letras trece" };
    if ($dig==4) {$letras= "$letras catorce" };
    if ($dig==5) {$letras= "$letras quince" };
```

```
    if ($dig==6) {$letras= "$letras dieciseis" };
    if ($dig==7) {$letras= "$letras diecisiete" };
    if ($dig==8) {$letras= "$letras dieciocho" };
    if ($dig==9) {$letras= "$letras diecinueve" };


}elsif ($decenas==2) {
    if ($dig==0) {$letras= "$letras veinte"; }
    if ($dig==1) {$letras= "$letras veintiuno"; }
    if ($dig==2) {$letras= "$letras veintids"; }
    if ($dig==3) {$letras= "$letras veintitrs"; }
    if ($dig==4) {$letras= "$letras veinticuatro"; }
    if ($dig==5) {$letras= "$letras veinticinco"; }
    if ($dig==6) {$letras= "$letras veintiseis"; }
    if ($dig==7) {$letras= "$letras veintisiete"; }
    if ($dig==8) {$letras= "$letras veintiocho"; }
    if ($dig==9) {$letras= "$letras veintinueve"; }

}else {
    if ($dig==0) { $letras= "$letras "; }
    if ($dig==1) { $letras= "$letras uno"; }
    if ($dig==2) { $letras= "$letras dos"; }
    if ($dig==3) { $letras= "$letras tres"; }
    if ($dig==4) { $letras= "$letras cuatro"; }
    if ($dig==5) { $letras= "$letras cinco"; }
    if ($dig==6) { $letras= "$letras seis"; }
    if ($dig==7) { $letras= "$letras siete"; }
    if ($dig==8) { $letras= "$letras ocho"; }
    if ($dig==9) { $letras= "$letras nueve"; }

}
    }
    $long--;

  }
    return($letras);
}
```

Apéndice D

Programa que convierte de texto a fonemas

```
#!/usr/bin/perl
#-------------------------------------------------------------------------
# Program: txt2phn.exe
# Input arguments: vocab-file
# Author: Esmeralda Uraga
# Last modification: 30-oct-2000, 3-nov-2000
#
# Description:
# This program generates the phonemic pronunciation of each word
# in the vocab file (a list of words in uppercases) applying a set
# of phonological rules and exceptions rules.
#-------------------------------------------------------------------------

#Grapheme to phone rules

while(<>) {

 s/\n//g;  #eliminating newline
 s/\s+//g; #eliminating  spaces
 $WORD=$_;
 $IN_WORD=$WORD;
 s/\'//g;  #eliminating isolated accents
 s/U:/g;
 s/N~//g;

 if ($WORD eq "AH") {
     $_=$OUT_WORD="ah";
 }elsif ($WORD eq "AY") {
```

```
        $_=$OUT_WORD="ay";
    }elsif ($WORD eq "EAH") {
        $_=$OUT_WORD="eah";
    }elsif ($WORD eq "EH") {
        $_=$OUT_WORD="eh";
    }elsif ($WORD eq "EM") {
        $_=$OUT_WORD="em";
    }elsif ($WORD eq "MM") {
        $_=$OUT_WORD="mm";
    }elsif ($WORD eq "PAU") {
        $_= "sil"; $OUT_WORD="";
    }elsif ($WORD eq "SIL") {
        $_= "sil"; $OUT_WORD="";
    }elsif (!($WORD=~/\!/)) {

        s/^/#/;
        s/$/#/;
        s/ф//g;
        s/Y#/i#/g;
        s/Z/S/g;
        s/CE/ sE/g;
        s/C/ s/g;
        s/CI/ sI/g;
        s/C/ s/g;
        s/CH/ ch /g;
        s/C/k /g;
        s/QU/k /g;
        s/X/KS/g;
        s/Q/k /g;
        s/K/k /g;
        s/k/kc k/g;
        s/D/D /g;
        s/V/V/g;
        s/B/V/g;
        s/#D/# dc d /g;
        s/LD/L dc d /g;
        s/ND/N dc d /g;
        s/#Y/dZc dZ/g;
        s/J/ x /g;
        s/#R/#r/g;
        s/RR/r/g;
#       s/ru/r u/g;
        s/LR/L r /g;
        s/NR/N r /g;
        s/SR/S r /g;
```

```
s/L/l/g;
s/ll/L/g;

s/#V/#bc b/g;
s/M/m/g;
s/mV/ m bc b/g;
s/NV/ m bc b/g;
s/Vm/bc b m/g;
s/VR/bc b R/g;
s/Vl/bc b l /g;
s/GE/ x E/g;
s/GI/ x I/g;
s/G/ x /g;
s/G/ x /g;
s/GUE/ G E/g;
s/GUI/ G I/g;
s/GU/ G /g;
s/GU/ G /g;
s/G/ G W/g;
s/N/n/g;
s/nk/N k/g;
s/nG/N gc g/g;
s/Gr/gc g r /g;
s/Gl/gc g l /g;
s/#G/gc g /g;

s/IA/ j a /g;
s/IE/ j e /g;
s/IO/ j o /g;
s/IU/ j u /g;
s/I/ j a /g;
s/I/ j e /g;
s/I/ j o /g;
s/I/ j u /g;

s/UA/ w a /g;
s/UE/ w e /g;
s/UI/ w i /g;
s/UO/ w o /g;
s/U/ w a /g;
s/U/ w e /g;
s/U/ w i /g;
s/U/ w o /g;

s// a /g;
```

```
s// e /g;
s// i /g;
s// o /g;
s// u /g;
s/A/ a /g;
s/E/ e /g;
s/I/ i /g;
s/O/ o /g;
s/U/ u /g;
s/W/ w /g;
s/R/ 3r /g;

s/F/ f /g;
s/H//g;
s// nj /g;
s/P/ pc p /g;
s/S/ s /g;
s/T/ tc t /g;
s/ch/ tSc tS /g;
s/l/ l /g;
s/m/ m /g;
s/Y/ dZc dZ /g;
s/3r/R/g;
s/r/RR/g;
s/#//g;
s/\s+/ /g;
($OUT_WORD=$WORD) =~ tr/A-Z/a-z/;  #to lowercases

}elsif  ($WORD=~/\!/) { # exception
    if (/RUIDO/) {
$_=ruido;
$OUT_WORD="<ruido>";
    }elsif ($WORD=~/NOVOCAL/) {
$_=nv;
$OUT_WORD="<novocal>";
    }elsif ($WORD=~/ASP/) {
$_=asp;
$OUT_WORD="<asp>";
    }else {
$_=sil;
$OUT_WORD="";
    }
}
```

```
$TAMANO=length($IN_WORD);
if ($TAMANO<4) {
     # create a list of phoneme-words
     push(@phonemes,"$IN_WORD \t\t\t \[$OUT_WORD\] \t\t\t $_");
  }elsif ($TAMANO<7){
      push(@phonemes,"$IN_WORD \t\t\t \[$OUT_WORD\] \t\t $_");
  }elsif ($TAMANO<12){
      push(@phonemes,"$IN_WORD \t\t \[$OUT_WORD\] \t\t $_");
  }else{
      push(@phonemes,"$IN_WORD \t\t \[$OUT_WORD\] \t $_");
  }
}

print join("\n",@phonemes);
print "\n";
#s///g;

# las guardamos para cuando este la silabificacion:
## s/N+V/M+B/g;
## s/#R/RR/g;
## s/(N|L|B)+R/(N|L|B)+RR/g;
```