UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
COORDINACION DEL DOCTORADO EN CIENCIAS APLICADAS

TRATAMIENTO Y RECONOCIMIENTO AUTOMATICO DE SEÑALES DE LA VOZ VENEZOLANA

Autor: Ing., Msc., José Luciano Maldonado
Tutor: Ing., Msc., PhD., Eliezer Colina Morles

Trabajo de tesis presentado ante la Ilustre UNIVERSIDAD DE LOS ANDES para obtener el grado de DOCTOR EN CIENCIAS APLICADAS de la Facultad de Ingeniería

Mérida, Venezuela, 2003

AGRADECIMIENTO

Al Instituto de Estadística Aplicada y Computación (IEAC), FACES, ULA, a través de sus directores.

A Eliezer Colina (Facultad de Ingeniería), Elsa Mora (Facultad de Humanidades y Educación) y Manuel Rodríguez (Facultad de Ingeniería), todos respetados profesores titulares de la Universidad de Los Andes, quienes me brindaron sus conocimientos, cada uno en sus áreas respectivas. Pero sobre todo, por brindarme su confianza y apoyo moral.

A los doctores Antonio Bonafonte, Asunción Moreno y José Mariño, todos profesores del Grupo de Procesado del Habla del Departamento de Teoría de la Señal y Comunicaciones, de la Universidad Politécnica de Cataluña, Barcelona, España, quienes me brindaron sus conocimientos y su confianza, a la distancia vía correo electrónico, así como cuando estuve de pasantía en esa Universidad y cuando ellos estuvieron aquí.

Al doctor Antonio Cardenal, profesor del Departamento de Teoría de la señal y Comunicaciones de la Universidad de Vigo, España, por su confianza y por compartir conmigo sus conocimientos cuando estuvo en nuestra Universidad.

A los estudiantes de doctorado (hoy doctores) en Tecnología del Habla de la Universidad Politécnica de Cataluña, Barcelona, España, Pau Pachés, Maho y Jaume Padrell (a quién no conozco personalmente), por compartir sus conocimientos conmigo, y por ayudarme a solucionar problemas en mis experimentos, cuando estuve en esa Universidad y posteriormente, aquí vía correo electrónico.

A los doctores José Carlos Segura y Antonio Rubio, profesores de la Universidad de Granada, España, por facilitarme importante material relacionado con los trabajos sobre procesamiento de voz que han desarrollado en dicha Universidad y por la confianza que me brindaron en mi corta estadía allí.

El desarrollo de este trabajo generó las siguientes publicaciones, asistencias a congresos, charlas, tutorías de tesis de pregrado, entre otras actividades:

- Maldonado J. L., Automatic Voice Recognition: Gender and Dialect using Venezuelan Speech. Revista Internacional, Información Tecnológica del Centro de Información Tecnológica de Chile, Vol. 13 Nro. 4, 2002.
- Maldonado J. L., Reconocimiento Automático del Habla Venezolana: Resultados de pruebas realizadas con frases pronunciadas por mujeres. Revista Técnica de Ingeniería, Universidad del Zulia, Venezuela, Vol. 25, Nro. 3, 2002.
- Maldonado J. L., Sistema Incremental Generador de Oraciones y de Descodificación Lingüística. XVI Congreso de la Sociedad Española para el Procesamiento del lenguaje Natural, SEPLN2002, Vigo, España, 2000.
- Maldonado J. L., Sistema de Descodificación Lingüística para Reconocedores Automáticos de la Voz. VI coloquio Venezolano de Bioingeniería, San Cristóbal, Venezuela, 2000.
- Maldonado J. L., Automatic Sentence Generator System for Speech Recognition Applications. International Conference on Modeling, Simulation and Neural Networks, MSNN-2000, Mérida-Venezuela, 2000.
- Maldonado J. L., Construcción de Sistemas de Reconocimiento Automático del Habla. Primer Seminario Binacional de Ingeniería de Sistemas, Universidad Francisco de Paula Santander, Colombia, 2001. Seminario invitado.
- Maldonado J. L., Resultados del reconocimiento Automático del habla Venezolana. Un caso de aplicación de los Modelos Ocultos de Markov. I Jornadas de Ingeniería Informática, UNET, San Cristóbal-Venezuela, 2001. Seminario invitado.
- Maldonado J. L., Resultados del reconocimiento Automático del habla Venezolana. IV Jornadas de Estudios Estadísticos, ULA, Mérida-Venezuela, 2001.
- Maldonado J. L., La Tecnología del Habla en la Universidad de Los Andes. I Jornadas de Informática, IUT, San Cristóbal-Venezuela, 2001. Seminario invitado.
- Maldonado J. L., La estadística como herramienta para el desarrollo de sistemas automáticos reconocedores del habla. Revista Economía Nueva Etapa de la Facultad de

Ciencias Económicas y Sociales de la Universidad de Los Andes, Mérida, Venezuela, No. 14, 1998.

- Maldonado J. L., La estadística como herramienta para el desarrollo de sistemas automáticos reconocedores del habla. I Jornadas de Estudios Estadísticos, FACES, ULA, 1998.
- Maldonado J. L., Construcción de Sistemas reconocedores del Habla a través de Modelos Ocultos de Markov. IV Jornadas Científico Técnicas de la Facultad de Ingeniería, ULA, 1998.
- Maldonado J. L., Un reconocedor automático de los fonemas Vocálicos basado en una Red Neural. IV Jornadas Científico Técnicas de la Facultad de Ingeniería, ULA, 1998.
- Maldonado J. L., Algoritmos para el desarrollo de reconocedores del Habla. Seminario dictado en el Postgrado de Ingeniería de Control, ULA, 1998.
- Maldonado J. L., Construcción de Sistemas Automáticos reconocedores del Habla. Seminario dictado a estudiantes de los últimos semestres de Ingeniería Electricista, ULA, 1998.
- Maldonado J. L., Un ejemplo de construcción de un sistema automático para reconocimiento de palabras (resultados prácticos). Seminario dictado en el Postgrado de Ingeniería de Control, ULA, 1998.
- Maldonado J. L., El estado del arte de la Tecnología del Habla. Seminario dictado en el Postgrado de Ingeniería de Control, ULA, 1998.

- **Tutorías de tesis:**

1. Andueza L., Diseño de un software para controlar procesos mediante el reconocimiento de voz a través de INTERNET. Ingeniero Electrónico, UNET, San Cristóbal-Venezuela. En ejecución.

2. Dubourdieu M., Distinción de Dialectos en forma automática. Ingeniero electricista, ULA, Mérida-Venezuela, 2002.

3. Moreno Y., Manejo automático de aparatos domésticos con comandos orales. Ingeniero electricista, ULA, Mérida-Venezuela, 2002.

4. Cumana J., Reconocimiento de voz con Modelos Marcovianos Ocultos: Dígitos Contínuos. Ingeniero Electricista, ULA, Mérida-Venezuela, 2001.

5. Paredes E., Evaluación del perceptrón multicapa para reconocimiento de voz. Ingeniero Electricista, ULA, Mérida-Venezuela, 1998.

- **Otros:**

1. Pasantía en la UPC, Barcelona, España, donde se realizaron pruebas de reconocimiento del español Peninsular y los primeros contactos con el HTK.

2. Participación en la construcción de la base de datos SPEECHDAT Venezolana 2000, obtenida por medio de telefonía fija (primera y única, hasta ahora, base de datos de la voz venezolana para fines de reconocimiento automático).

3. Maldonado J. L., Construcción de modelos del habla venezolana para pruebas de reconocimiento automático de oraciones. Trabajo aceptado en el EIGHT INTERNATIONAL SYMPOSIUM ON SOCIAL COMMUNICATION, Cuba enero 2003, sección de reconocimiento automático del habla.

4. Maldonado J. L., Redes Neuronales Artificiales. Curso dictado en la Universidad de Granada, España (marzo del 2003), en el marco de la Red Temática sobre Tecnologías del Habla, que mantiene la ULA con esa y otras universidades españolas.

5. Miembro del jurado evaluador del trabajo "Definición de parámetros para la descripción acústica de sistemas vocálicos", de la profesora Carmen Elena Contreras a través del cual, dicha profesora ascendió a la categoría de profesor Agregado, (marzo 2003).

6. Actualmente se coordina la construcción de una segunda base de datos SPEECHDAT Venezolana , obtenida por medio de telefonía celular.

DEDICATORIA

Este trabajo está dedicado a mi familia toda,
en particular a mi madre a quién el destino
le negó la oportunidad de verme crecer como profesional.

A mi esposa Blanca Elena.

Y muy particularmente a mis hijas **Luz Elena y Malvy Elena**,
a quienes culpo de no dejarme trabajar con más dedicación,
pero la verdad es que disfruto mucho acompañándolas en sus diversas
competencias regionales y nacionales de baloncesto.

A nuestra ULA por permitirme formar parte de ella.

RESUMEN

Este trabajo tuvo como objetivo hacer reconocimiento automático de unidades del español hablado en Venezuela. Las actividades realizadas consistieron en: obtención de las señales de voz, preprocesamiento de dichas señales, selección y construcción de los modelos de voz, construcción por programación de diversos reconocedores y pruebas de reconocimiento.

Un objetivo subyacente del trabajo consistió en contribuir al desarrollo de la tecnología del habla en Venezuela, específicamente con el uso de la teoría de los Modelos Ocultos de Markov, que es la herramienta base de la mayoría de los sistemas de reconocimiento comerciales actuales, adoptada hace aproximadamente veinticinco años a este campo y que tiene aplicaciones en diversas áreas, pero que en este País es poco conocida.

Las actividades llevadas a cabo, estuvieron orientadas a la búsqueda y construcción de una base de datos del habla venezolana para fines de su reconocimiento automático y al desarrollo experimental de elementos básicos de los sistemas de reconocimiento para aplicaciones de tiempo real, cuya comunicación hablada se realice a través de líneas telefónicas fijas, en forma independiente del hablante y donde el lenguaje sea el español hablado por los venezolanos. En ese sentido, se crearon modelos de la voz de mujeres, modelos de la voz de hombres, modelos híbridos de la voz de hombres y de mujeres, en los diferentes casos se trabajó con unidades de fonos y de palabras que permitieron reconocer pronunciaciones de palabras conectadas, oraciones, el género del hablante y el dialecto del hablante entre varios dialectos venezolanos.

Los resultados experimentales que se obtuvieron, así como la base de datos con la que se cuenta actualmente, muestran que a partir de ahora se puede modelar en forma satisfactoria el habla venezolana para diversas aplicaciones de reconocimiento donde la entrada sea a través de líneas telefónicas fijas. Un escenario real de prueba será aquel donde se aprovechen los resultados de este trabajo para construir un prototipo reconocedor autónomo del habla de los venezolanos.

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

CAPITULO I

INTRODUCCION

Desde que el hombre tuvo conciencia de que pertenece a un ambiente donde se interactúa con seres de su misma especie y de otras, la comunicación hablada fue y seguirá siendo la forma más destacable de intercambio de información entre los humanos. De hecho, la comunicación por medio de la palabra hablada ha trascendido, producto del avance tecnológico, su forma natural y se ha extendido a través de medios como el teléfono en sus distintas modalidades (fijos y celulares), el cine, la radio, la televisión, las videoconferencias y más recientemente por Internet.

En la actualidad cobra aun mayor importancia la comunicación hablada, puesto que se está constituyendo en la forma en que las personas desean comunicarse con las máquinas. Es decir, está planteado que las máquinas adquieran habilidades fundamentales del hombre como son hablar, escuchar, analizar y aprender; y por esa razón es que han venido apareciendo desde hace algún tiempo, dispositivos y pequeños aparatos hasta computadores propiamente dichos, que cumplen algunas tareas muy específicas en las que son capaces de recibir y entregar información por medio de la voz.

Lo anterior expresado no significa que el reto de establecer comunicación hablada con las máquinas haya sido superado, sólo que se han logrado avances y resultados importantes, pero que en esa dirección queda mucho por realizar cuando se persigue la construcción de sistemas con los que se pueda comunicar cualquier persona, en cualquier ambiente y en cualquier idioma.

De hecho, desde los primeros trabajos sobre reconocimiento de dígitos realizados en los años cincuenta del siglo pasado hasta los sistemas actuales de dictado, transcripción y diálogo, ha habido una gran actividad de investigación relacionada con el reconocimiento automático y la producción del habla. Todo este esfuerzo de investigación ha dado lugar a la publicación de miles de artículos en revistas especializadas y en las secciones de los congresos más

importantes de procesamiento de las señales, de inteligencia artificial, de tecnología del habla y de procesamiento del lenguaje natural.

En las últimas cinco décadas de investigación han ocurrido planteamientos de tareas relacionadas con la tecnología del habla, cada vez más ambiciosas que se han ido solucionando en la medida que la tecnología del momento lo permite. De forma progresiva, el reconocimiento de voz ha desarrollado su capacidad desde las primeras aplicaciones de palabras aisladas, limitadas a un único usuario y con un vocabulario muy restringido hasta sistemas independientes del hablante y capaces de reconocer habla continua con vocabularios de varios miles y hasta decenas de miles de palabras.

En el mundo actual de la investigación en tecnología del habla, las tareas que se están acometiendo son altamente complejas, puesto que la interacción usuario-máquina se busca que sea natural, que se puedan establecer verdaderos diálogos y donde pueda transcribirse de forma eficiente, por ejemplo, conversaciones telefónicas, y hacer consultas a bases de datos con habla espontánea.

La tecnología que ha permitido alcanzar los niveles actuales no es exactamente nueva. La base de la mayoría de los reconocedores actuales son los Modelos Ocultos de Markov, cuya aplicación inicial en este campo apareció hace aproximadamente veinticinco años. Los avances observados en los últimos tiempos son debidos en gran medida al continuo desarrollo de la electrónica y con ella, la capacidad de los computadores, y por supuesto, a la aparición y mejora de otras herramientas algorítmicas complementarias que permiten lograr que todos los elementos involucrados en el proceso de reconocimiento sean más eficientes.

En este sentido hay continuos avances que se pueden observar en el modelado acústico, en el uso de modelos con contexto y libres de contexto, en las técnicas de entrenamiento, en la adopción de técnicas no tradicionales del área (ANN y lógica difusa, por ejemplo), en los métodos de parametrización de las señales, en la construcción de bases de datos de voz, en las técnicas de modelado de lenguaje, en la imposición de restricciones gramaticales en el reconocimiento, etc. Todo estas, son áreas de investigación dentro del campo de la tecnología

del habla que permanecen activan y que han conducido a mejorar cada día las tasas de reconocimiento [44][48][50][54][55][56][57][58][59][60].

1.1. MOTIVACIÓN PARA LA REALIZACIÓN DEL TRABAJO

Uno de los subcampos de investigación de la Tecnología del Habla es el Reconocimiento Automático, que tiene como objetivo darle a la máquina la capacidad "de escuchar", es decir, identificar la secuencia de unidades del lenguaje hablado llámense fonos, sílabas, palabras y hasta oraciones presentes en una pronunciación dada como entrada, almacenar dicha secuencia y eventualmente realizar alguna acción dependiendo del mensaje detectado en la secuencia. Este subcampo tiene una gran variedad de aplicaciones en la vida diaria y sin embargo, en Venezuela este tipo de tecnología ha sido muy poco explorada.

Con la idea de abordar este subcampo de investigación y contribuir en algún grado con su desarrollo en Venezuela se planteó la realización de una serie de experimentos que consistieron, en general, en la construcción de una base de datos del habla venezolana para fines de reconocimiento, en la parametrización de dichas señales, en la construcción de modelos matemáticos del español venezolano, y en la realización de pruebas de reconocimiento del habla venezolana, con el objetivo de crear la posibilidad cierta de que en el mediano y largo plazo se pueda construir algún sistema de reconocimiento que trate con el habla propia de los venezolanos.

Este tipo de trabajos cobra importancia cuando se toma en cuenta que diversos centros de investigación a nivel mundial, han enfocado sus desarrollos a incorporar poco a poco modelos de voz de unas cuantas personas, con pronunciaciones de un tema particular y en un ambiente también particular con la finalidad de que en forma incremental, se pueda llegar en el futuro a sistemas de reconocimiento más generales. A manera de ejemplo, se puede encontrar en la literatura [42][45][51][54][61], trabajos relacionados con el desarrollo de base de datos del gallego, del mandarín, del japonés, del inglés americano, del inglés británico, del español peninsular, etc., para fines de reconocimiento automático. De la misma manera, en cada caso se busca dependiendo del país, del idioma y hasta de la cultura, y para aplicaciones particulares, desarrollar reconocedores de la voz.

1.2. CONTRIBUCIONES DEL TRABAJO

Las contribuciones de esta tesis al desarrollo de la Tecnología del Habla son las siguientes:

- Se dispone ahora de la base de datos SpeechDat Venezolana producto de la colaboración de la Universidad de Los Andes de Venezuela con la Universidad Politécnica de Cataluña de España. Esta base de datos es la primera orientada al reconocimiento automático de la voz que se concibe en Venezuela y permitirá que la voz venezolana esté disponible para ser incorporada en cualquier sistema de reconocimiento que se desarrolle en cualquier centro de investigación del mundo.

- La producción de diversos tipos de modelos matemáticos de la voz del español hablado en Venezuela, lo cual permitió y seguirá permitiendo hacer reconocimiento automático del habla de los venezolanos. Este tipo de investigación es inédita en este País y a nivel mundial no se ha experimentado con habla venezolana. Los modelos de la voz que se crearon fueron: modelos de fonos de la voz de mujeres, modelos de fonos de la voz de hombres, modelos de fonos híbridos de la voz tanto de mujeres como de hombres y al mismo nivel, modelos de palabras.

- La generación de un modelo de la gramática de fechas con la particularidad de que se trata de un modelo específico de las formas en qué los venezolanos pronuncian esas unidades del lenguaje. Este modelo gramatical permitió hacer reconocimiento automático de oraciones, donde las pronunciaciones de entrada al reconocedor eran fechas.

- La construcción por programación de un tipo de reconocedor que permitía distinguir si la voz que se daba como entrada pertenecía a un hombre o a una mujer. Para ello se emplearon los modelos separados del habla de mujeres y los modelos del habla de hombres. Este hecho tiene singular relevancia debido a que para muchas aplicaciones prácticas reales, los modelos separados constituyen una alternativa para superar el rendimiento que presentan los modelos híbridos.

- La realización de reconocimiento automático de estructuras complejas del habla como son las oraciones, donde el hablante puede ser cualquier persona de cualquier parte del territorio venezolano cuya comunicación hablada la realice a través del idioma español. Para este tipo de reconocimiento se utilizaron los modelos híbridos y la gramática de fechas venezolana.

- La implementación de dos formas de hacer reconocimiento automático de dialectos venezolanos. Se dividió el territorio venezolano en cinco zonas dialectales, para cada zona se crearon modelos de la voz. La tarea del reconocedor era identificar a qué zona dialectal pertenecía el hablante cuya voz se daba como entrada. La importancia de estos experimentos es que estuvieron orientados al reconocimiento de dialectos, en particular dialectos venezolanos, un campo no investigado en el mundo puesto que la orientación general es hacia la identificación o reconocimiento de los idiomas y no al de los dialectos.

- Se reporta un algoritmo que permite elaborar modelos de lenguaje. En la actualidad, en el campo de los reconocedores, la atención se centra más en el desarrollo del módulo de descodificación acústico que en el módulo de descodificación lingüístico, y es claro que del rendimiento de éste último depende en gran medida el rendimiento del reconocedor como un todo, es por esa razón que se propuso y se programó un algoritmo como una alternativa a las existentes para contribuir al desarrollo de descodificadores lingüísticos para sistemas de reconocimiento de voz y/o para sistemas automáticos generadores de oraciones.

- Se determinó que las redes neurales perceptrónicas multicapas entrenadas con BKP no son adecuadas para hacer reconocimiento de voz a menos que se realice un preprocesamiento riguroso de las señales de entrada. Esto se pudo corroborar en un caso de reconocimiento de pronunciaciones de palabras aisladas, independientes del hablante donde por un lado se hizo uso de las técnicas de los Modelos Ocultos de Markov de observaciones discretas y por otro lado, se utilizó el tipo de redes mencionado.

- El reconocimiento que se realizó fue independiente del hablante. Los modelos de la voz que se construyeron y todas las pruebas de reconocimiento que se realizaron estuvieron orientados a la construcción de sistemas robustos, en el sentido de que en los diferentes casos, se aceptaban pronunciaciones de múltiples hablantes venezolanos, cuyas edades oscilaban entre los 18 y los 60 años. Este tipo de pruebas están revestidas de una particular importancia debido a que la mayoría de los sistemas comerciales que funcionan en la actualidad solo trabajan con la voz de una persona o de un grupo reducido de personas.

- En general, se reportan resultados de hacer reconocimiento de fonos, palabras y oraciones del habla de los venezolanos donde se trabaja con voz con calidad telefónica fija y que pueden tener aplicaciones de tiempo real.

La contribución más importante, se refiere al hecho de que este trabajo constituye el primer intento de envergadura por hacer reconocimiento automático del español venezolano, el cual dió lugar a las primeras publicaciones relacionadas con la Tecnología del habla en Venezuela, que seguramente se constituirá en un punto de partida para adoptar este tipo de tecnología en el País.

1.3. PROCEDIMIENTO EXPERIMENTAL DEL TRABAJO

El desarrollo del trabajo comprendió tres grandes etapas: la primera estuvo relacionada con la revisión del proceso de construcción de los sistemas de reconocimiento automático del habla, la segunda, estuvo relacionada con la definición del dominio de las pronunciaciones sobre las que se realizó el reconocimiento y una tercera etapa que comprendió la construcción de sistemas a través de los cuales se realizaron las pruebas de reconocimiento.

1.3.1. PRIMERA ETAPA

Del estudio del proceso de construcción de los sistemas de reconocimiento automático del habla, se pudo determinar que dichos sistemas están constituidos de una cantidad de módulos implementados en serie y en paralelo a través de diversas herramientas algorítmicas y

matemáticas complejas que están en constante evolución, con el fin de lograr cada día tasas de reconocimiento más altas. Dentro de ese conjunto de herramientas de avanzada tecnológica, la que menos ha evolucionado y que sin embargo constituye la base sobre la cual están construidos los modelos de la voz de la mayoría de los reconocedores comerciales actuales, es la que comprende los Modelos Ocultos de Markov [1][15][17][25][56].

Como efectivamente, los algoritmos basados en los Modelos Ocultos de Markov constituyen la herramienta que ha prevalecido y que prevalecerá por años en el campo del reconocimiento (aun cuando evolucione o se mezcle, como está ocurriendo, con otro tipos de tecnologías como las ANN o la Lógica Difusa), es imprescindible que toda persona que trabaje en procesamiento de voz desde el punto de vista ingenieríl, realice una revisión minuciosa del funcionamiento y aplicabilidad de esta tecnología sin descuidar por supuesto otras tecnologías emergentes. En función de ésto, la primera actividad experimental relacionada con el trabajo descrito en esta tesis, consistió en la implementación de todos los algoritmos propios de la Teoría de los Modelos Ocultos de Markov (MOM), a través de programación en lenguaje C para realizar reconocimiento de palabras aisladas. La programación de estos algoritmos tiene su importancia en el hecho de que teóricamente funcionan perfectamente, sin embargo, en la práctica su implementación tiene que estar acompañada (específicamente los algoritmos del tipo Baum-Welch) de algoritmos de escalado (para evitar pérdida de precisión en la representación numérica debido a la gran cantidad de multiplicaciones de números con valores entre 0 y 1) muy difíciles de construir. Con esta actividad se logró la construcción de un pequeño sistema que permitía construir modelos de voz a partir de señales de pronunciaciones parametrizadas y también hacer pruebas de reconocimiento de señales tanto de entrenamiento como de test.

Para esta primera etapa no se contaba con una base de datos del habla venezolana para construir los modelos y para realizar las pruebas de reconocimiento, por lo que se utilizó una pequeña base de datos facilitada por la Universidad Politécnica de Cataluña de España que consistía de pronunciaciones aisladas de los dígitos del Catalán. Vale la pena tomar en cuenta que es imposible hacer pruebas de reconocimiento del habla en un idioma particular, si no se cuenta con una base de datos de ese idioma para hacer el entrenamiento, y que sin embargo, la programación del sistema es independiente de la base de datos, en el sentido de que en un

momento se puede entrenar para hacer reconocimiento del habla en un idioma y posteriormente se puede entrenar para reconocer el habla de otro idioma.

Esta actividad, que será descrita en el capítulo 3, estuvo acompañada de la programación de un sistema de reconocimiento basado en Redes Neurales Perceptrónicas entrenadas con el algoritmo BKP [18][19][20][27][36][37], como una forma de comparar si este tipo de redes podían competir en eficiencia en el reconocimiento de palabras aisladas con los algoritmos basados en MOM.

1.3.2. SEGUNDA ETAPA

En esta etapa, el trabajo estuvo orientado a la búsqueda de una base de datos de voz del español hablado en Venezuela que permitiera cumplir uno de los objetivos de la tesis como era realizar reconocimiento automático del habla de los venezolanos. La construcción de dicha base de datos era otro de los objetivos más importantes de este trabajo.

La construcción de bases de datos de voz (corpora o corpus cómo se les conoce en la tecnología del habla) ha despertado tal interés en los investigadores del campo del reconocimiento, puesto que constituyen la materia prima más importante para el desarrollo y la aplicabilidad de los reconocedores, que en la actualidad existen diversas bases de datos que coleccionan la voz del inglés, el español peninsular, el catalán, el gallego, el japonés, el francés, el alemán, etc., y hasta existen formatos estándares para dicha construcción como es el SPEECHDAT [46][47][51][61] y un comité internacional para la estandarización de las bases de datos de voz (International Committee for the Co-ordination and Standarisation of Speech Databases and Assesment Techniques, COCOSDA)

El proceso de búsqueda de una base de datos de esta naturaleza es muy complicado, puesto que se requiere mucho material humano y económico aparte de los medios electrónicos que transportan y almacenan la voz. En primera instancia, en este trabajo, se pensó construir una base de datos a través de micrófono, sin embargo, afortunadamente la Universidad Politécnica de Cataluña de España, tenía planteado un proyecto (el Proyecto SALA I, Speech Accros Latinoamérica [32][51]) para construir una base de datos del español hablado en

Latinoamérica a través de telefonía fija. La Universidad de Los Andes, bajo el soporte económico y la orientación de la Universidad Politécnica de Cataluña llevó a cabo la construcción de la Base de Datos SPEECHDAT Venezolana. De esta manera, actualmente se cuenta con una base de datos de voz venezolana dentro del conjunto de bases de datos tipo SPEECHDAT que existen en el mundo y que está disponible para la investigación en el campo del reconocimiento y de la lingüística, no solo en Venezuela sino para cualquier centro de investigación y desarrollo.

En el capitulo 4 se muestran detalles de la construcción de la base de datos del español hablado en Venezuela.

1.3.3. TERCERA ETAPA

Las actividades cubiertas en esta etapa estuvieron orientadas a la creación de modelos de unidades del habla del español hablado en Venezuela y a su reconocimiento automático. Para ello se hizo uso de la base de datos SPEECHDAT Venezolana.

Entrenamiento o creación de los modelos de voz

El proceso de construcción de los modelos se inició con la selección de las unidades a modelar, las cuales fueron: fonos y palabras del habla de los venezolanos (del habla de mujeres, del habla de hombres, del habla de mujeres y de hombres a la vez, por zonas dialectales y de todo el territorio nacional).

Para la construcción de los modelos se utilizó el sistema de desarrollo HTK [4] que permite construir Modelos Ocultos de Markov de la voz a través del algoritmo Baum-Welch [1] [6][8][13][14]. Un paso previo a la construcción de los modelos consistió en la segmentación y transcripción ortográfica de los sonidos del habla (fonos y palabras) presentes en cada archivo de voz. De la misma manera se realizó la parametrización de las señales a coeficientes cepstrales más la energía y las primeras y segundas derivadas por segmentos de 25 milisegundos desplazados cada 10 milisegundos. Para crear cada modelo, finalmente se utilizaron conjuntos de patrones parametrizados de cada fono o palabra según el caso.

Reconocimiento de voz

Posterior a la construcción de los modelos de los fonos y de las palabras, se construyeron a través de HTK y del algoritmo Viterbi [1][6][8][13][14], programas que implementan un reconocedor de pronunciaciones de dígitos conectados con modelos de palabras, un reconocedor de pronunciaciones de dígitos conectados con modelos de fonos, dos reconocedores que determinan el dialecto venezolano del hablante: uno que utiliza modelos de palabras y otro que utiliza modelos de fonos, un reconocedor que determina que la voz que se da como entrada pertenece a un hombre o a una mujer, un reconocedor de oraciones con fonos de mujeres, un reconocedor de oraciones con fonos de hombres, un reconocedor de oraciones con fonos híbridos de la voz de hombres y de mujeres.

Se elaboró una gramática de fechas que implementa todas las formas en que los venezolanos pronuncian esas unidades del habla. Esta gramática cobra importancia, cuando se desea un reconocedor de pronunciaciones de fechas, puesto que contribuiría al reconocimiento cuando su entrada sean fechas venezolanas (y tal vez de algunos paises de Latinoamérica), pero no de España, puesto que no considera expresiones del tipo "la una menos cuarto", por poner un ejemplo, que es una pronunciación propia de ese País.

Se escogió el reconocimiento de fechas para realizar las pruebas de reconocimiento de habla contínua debido en primer lugar, a que lo realmente importante era verificar si los modelos de los fonos venezolanos estaban bien construidos, y en segundo lugar, averiguar si se podían utilizar en el reconocimiento de algún tipo de oraciones. Por otro lado, se realizó reconocimiento de fechas debido a que el sistema de desarrollo HTK adolece de una forma de construcción de modelos de lenguaje para aplicaciones complejas, a pesar de que es la herramienta más popular dentro del mundo de la construcción de reconocedores.

Los reconocedores fueron construidos a través de HTK [4] sobre Linux, para ser utilizados a nivel de laboratorio, por lo tanto no son transportables directamente como sistemas que puedan ser utilizados en máquinas que trabajen bajo otro sistema operativo. Se quiere decir con lo anterior, que si desea construir algún sistema de reconocimiento que trabaje en forma autónoma y en tiempo real sobre un sistema operativo particular, se pueden usar los modelos

de voz creados y probados en este trabajo, y hacer la programación respectiva de la entrada por medio de una tarjeta de telefonía, programar tanto los algoritmos de reconocimiento utilizados aquí, como la presentación ortográfica de las entradas reconocidas. Esto es lo que permite precisamente HTK, probar y construir modelos y transportarlos para el desarrollo de sistemas de reconocimiento de aplicaciones específicas.

Las actividades desarrolladas en la tercera etapa se describen en los capítulos del 5 al 7.

1.3.4. MODELADO DEL LENGUAJE PARA RECONOCIMIENTO

Una actividad que también se llevó a cabo, pero que fue paralela a la segunda etapa y a buena parte de la tercera, tuvo que ver con la implementación por programación de un algoritmo que se propuso como una forma de realizar modelado de lenguaje, concretamente referido a modelar contextos de aplicaciones de reconocimiento. El modelado del lenguaje es un área abierta a la investigación, por lo que un punto de partida para diseñar e implementar descodificadores lingüísticos en Venezuela es precisamente esta forma a la que se le puede adaptar variantes de las técnicas más populares como son los n-gramas y el Backoff [22][33][56].

En el capitulo 11, se muestran los resultados de esta actividad.

1.4. ORGANIZACIÓN DE LA TESIS

El trabajo presenta por capítulos la siguiente distribución: el capítulo I, es el capítulo introductorio que se está describiendo, el capítulo II, en el cual se da una breve reseña teórica de qué son y cómo es la arquitectura de los sistemas de reconocimiento automático del habla; En el capítulo III, se describen las primeras pruebas de reconocimiento de voz que se realizaron en este trabajo, las cuales consistieron en la construcción de un reconocedor prototipo de los dígitos pronunciados en Catalán; en el capítulo IV, se describe la base de datos de voces venezolanas, que se utilizó en la mayoría de las pruebas; en el capítulo V, se presentan los resultados de un conjunto de pruebas de reconocimiento de secuencias de palabras conectadas, cuyas pronunciaciones corresponden a hombres y mujeres de Venezuela,

donde se trabajó con modelos de palabras; en el capítulo VI, se presentan los resultados de un conjunto de pruebas de reconocimiento de habla contínua con la voz de mujeres venezolanas, donde se trabajó con modelos de fonos; en el capítulo VII, se presentan los resultados de un conjunto de pruebas de reconocimiento de habla contínua con la voz de mujeres y hombres venezolanos, donde se trabajó también con modelos de fonos; en el capítulo VIII, se presentan los resultados de un conjunto de pruebas de reconocimiento de dialectos venezolanos y del género del hablante, donde se utilizaron modelos de palabras; en el capítulo IX, se presentan los resultados de un conjunto de pruebas de reconocimiento de dialectos venezolanos, donde se utilizaron modelos de fonos; en el capítulo X, se presentan los resultados de un conjunto de pruebas de reconocimiento de secuencias de palabras conectadas por medio de modelos de fonos; en el capítulo XI, se presenta un sistema que permite crear modelos de lenguaje de aplicaciones particulares, generar automáticamente oraciones a partir de esos modelos y hacer descodificación lingüística de secuencias de palabras.

Finalmente, se presentan las conclusiones y las recomendaciones producto del trabajo, la bibliografía que sirvió de soporte tanto a las pruebas, como a esta redacción, y los anexos necesarios para complementar la claridad de los experimentos y de los resultados obtenidos.

CAPITULO II

EL RECONOCIMIENTO AUTOMATICO DEL HABLA

2.1. INTRODUCCIÓN

En este capítulo se presenta la información referida a enmarcar el trabajo dentro del campo de la Tecnología del Habla, concretamente en el subcampo del reconocimiento; por esa razón, se explica de manera general, qué son los sistemas de reconocimiento automático de la voz, cómo es su arquitectura, sus características más relevantes, los diferentes tipos que existen y se señalan algunas de las herramientas más populares de construcción de tales sistemas.

2.2. LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Desde un punto de vista simplista, un sistema de reconocimiento automático de la voz, es cualquier máquina que tenga la habilidad de recibir mensajes entregados por usuarios humanos, en forma hablada.

Una definición más refinada es la que presenta Bonafonte A., [22], en su tesis doctoral: "un sistema de reconocimiento es un computador o cualquier máquina que pudiera aceptar comandos y preguntas que cualquier usuario formulase por medio de un micrófono o teléfono, entendiese el significado que se pretendía transmitir con esa frase y generase una respuesta". Una máquina con la robustez que implica esta última definición no existe todavía, es decir, no hay reconocedores que puedan recibir la voz de cualquier persona, interpretar en qué consiste el mensaje y en función de ello actuar y generar respuestas que incluso, podrían ser orales también.

El ambiente que se vive actualmente, respecto a los sistemas de reconocimiento, es el que se describe por medio de la figura 2.1. En esa figura, se puede observar que efectivamente se dispone de una máquina que recibe la señal acústica asociada a una frase pronunciada, que transforma esa señal acústica en una representación ortográfica perteneciente al idioma que

habla el usuario, y que finalmente almacena esa última representación, para su posterior uso. Este es el tipo de función que cumplen la mayoría de los reconocedores comerciales.

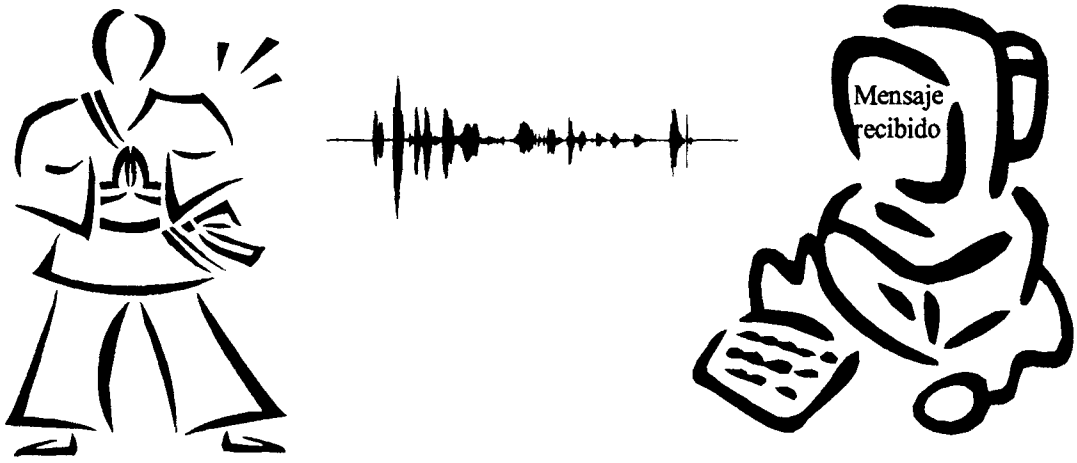


Figura 2.1. Sistemas de reconocimiento.

De cualquier manera, cada día predomina la tendencia a exigir algo más a esos sistemas. Entre esas exigencias se encuentran por ejemplo, las que se refieren a darles la capacidad de averiguar si las unidades del habla que se detectan en las señales acústicas, constituyen oraciones válidas de los lenguajes, y las que se refieren a que el sistema realice algún tipo de interpretación semántica de las oraciones, que pueda dar lugar a alguna operación particular que el usuario desee que el sistema ejecute, aparte de que se desea que reciban la voz de distintas personas y en idiomas distintos.

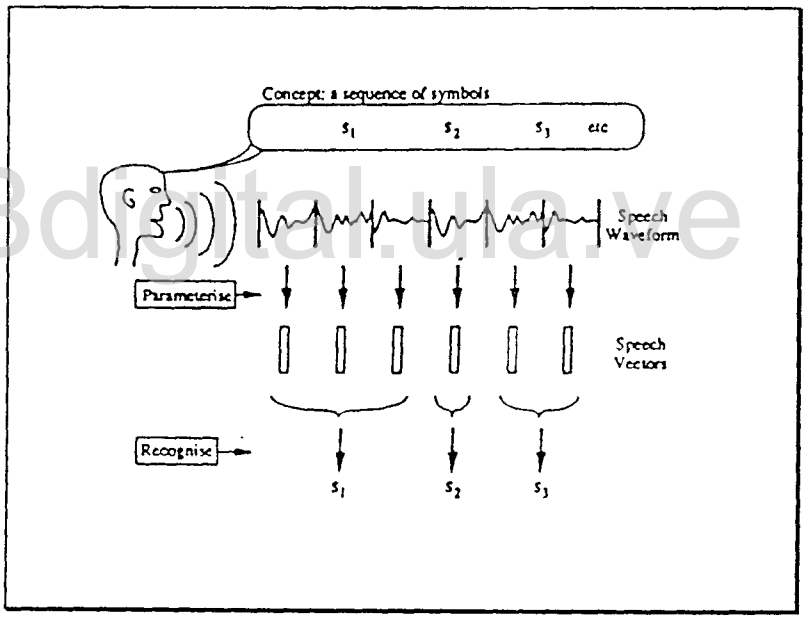
2.3. EL PROCESO DEL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

El proceso de reconocimiento, parte de una idea que el usuario desea transmitirle a la máquina. Para transformar esa idea en una forma que la máquina la pueda recibir, recurre al uso de un conjunto de reglas gramaticales propias del idioma que habla. Luego, de una manera que no está totalmente clara para nadie, el cerebro del usuario produce una serie de órdenes con el fin de hacer que se activen todos los mecanismos propios de su sistema de producción de voz, dando lugar al acto de hablar y en consecuencia, dando origen a la serie de sonidos que constituyen los elementos de las palabras que forman el mensaje [12]. Según Mora E., [67]

"las diferentes etapas cognitivas de la producción del habla son: una etapa de conceptualización, una de formulación y una de articulación".

Los sonidos asociados al mensaje viajan por el aire en forma de ondas y llegan a elementos transductores, que en algunos casos pueden ser micrófonos y en otros casos sistemas de telefonía, que los transmiten a la máquina reconocedora. En la máquina ocurre un proceso que consiste en tomar la señal digitalizada del mensaje, y en extraer de ésta, vectores de propiedades (vectores de parámetros) por segmentos.

En la figura 2.2 se muestra de manera gráfica, la forma general en que ocurre el proceso de reconocimiento.



(figura tomada del manual de HTK [4])

Figura 2.2. El proceso de reconocimiento automático de la voz

Después de contar con una secuencia de vectores de propiedades para la señal correspondiente al mensaje que está recibiendo, el reconocedor desencadena una estrategia de búsqueda, a través de la cual, partiendo de unos modelos acústicos que se encuentran almacenados en su memoria, logra asociar esa secuencia de vectores con algunos de esos modelos.

Previamente a todo proceso de reconocimiento, hay una etapa de construcción de los modelos acústicos, dichos modelos no son más que estructuras matemáticas que representan elementos de las palabras del lenguaje, a partir de un conjunto de vectores de propiedades (patrones) extraídos de muchas realizaciones acústicas de las palabras, y/o de realizaciones acústicas de elementos de las palabras.

Los sistemas de reconocimiento cuentan con mecanismos algorítmicos que determinan si alguna secuencia parcial de los vectores de propiedades acústicas del mensaje a reconocer, se corresponde con alguno de los modelos acústicos, y de esta manera, identifican cuáles elementos del lenguaje (fonos, sílabas, palabras, etc.) están presentes en ese mensaje, para posteriormente hacer una transcripción ortográfica (en grafemas) de todo el mensaje.

2.4. ARQUITECTURA GENERAL DE LOS SISTEMAS DE RECONOCIMIENTO

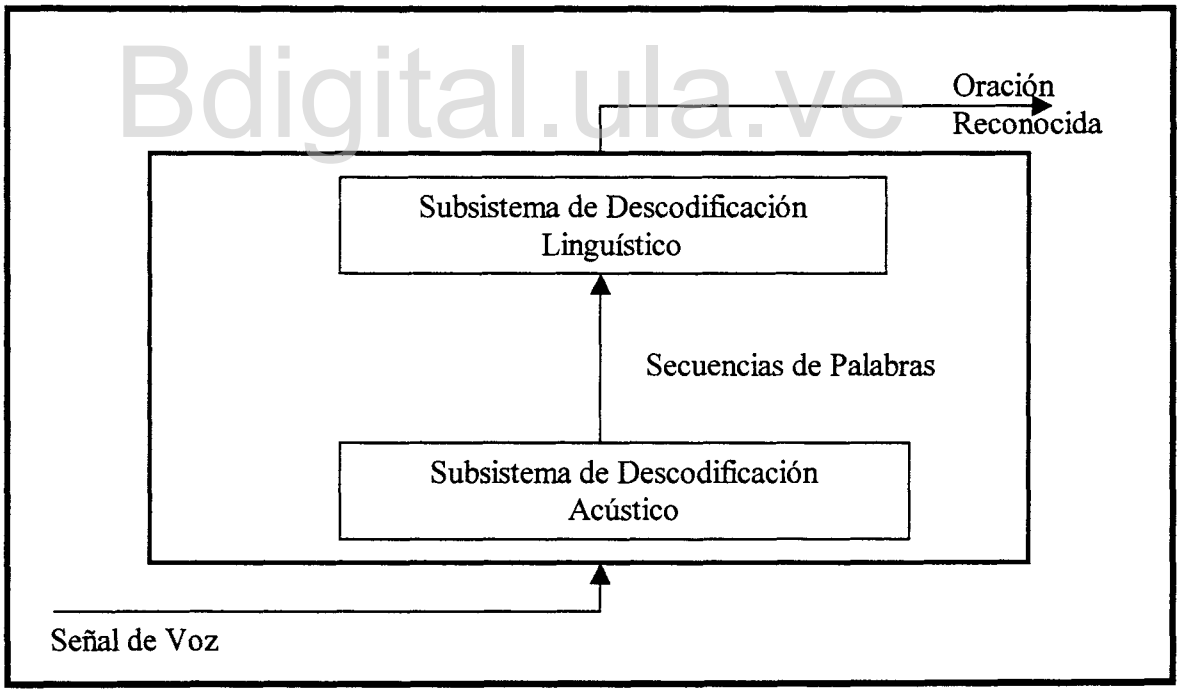


Figura 2.3. Arquitectura general de los Sistemas de Reconocimiento.

Un sistema de reconocimiento típico y realizable desde el punto de vista de su implementación práctica, comprende los componentes básicos que se muestran en la figura 2.3. Sin embargo, por años, se han manejado diversas arquitecturas, que por supuesto dependen del grado de

exigencia del diseñador, pero la mayoría coinciden en presentar como sus grandes componentes, un subsistema de descodificación acústico y un subsistema de descodificación lingüístico. Por esta razón, es que aquí se describen a los reconocedores como sistemas integrados por esos dos módulos, tomando en cuenta que desde el punto de vista descriptivo pueden existir gráficas donde se presente de manera más fina la relación de esos componentes, es decir, pueden aparecer arquitecturas donde se muestren esos componentes como integrados por otra serie de elementos, pero también puede darse el caso en que aparezcan los dos grandes módulos integrados en uno sólo.

2.4.1. El Subsistema de Descodificación Acústico

El módulo de descodificación acústico es el que se encarga de recibir la señal de la voz y de realizar la identificación inicial de los distintos sonidos del habla, presentes en ésta.

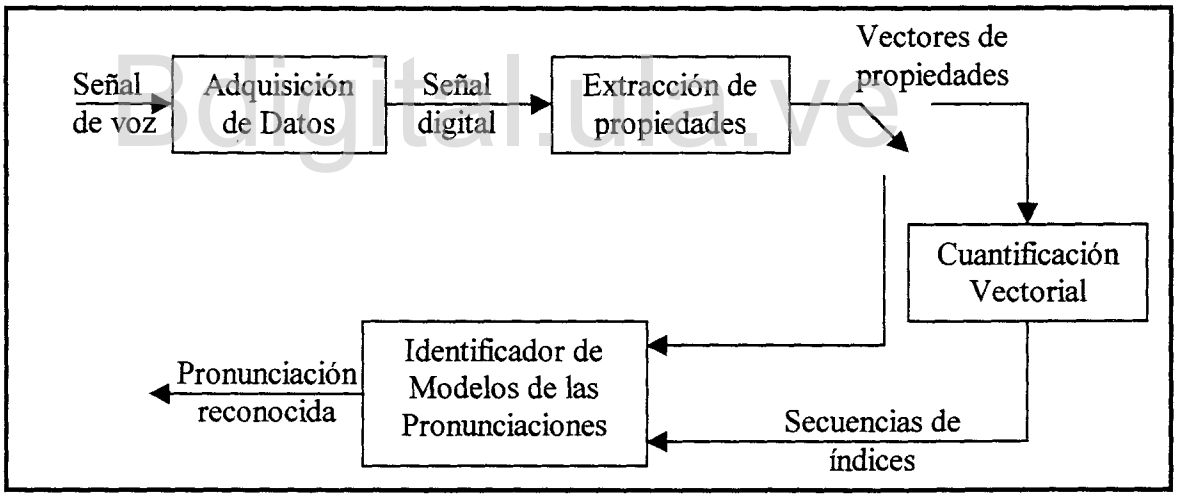


Figura 2.4. Sistema de descodificación acústico.

En base a la figura 2.4., se puede determinar que el subsistema de descodificación acústico realiza una serie de funciones que se resumen de la siguiente manera: se encarga de transformar la señal acústica del mensaje en una señal eléctrica inicialmente y luego, en una secuencia digital. Una vez que dispone de la representación digital del mensaje, activa una serie de algoritmos que se encargan de segmentar la señal y de realizar un análisis por tramos

de la señal. Este análisis consiste en tomar cada segmento y obtener de éste, un vector de parámetros.

La idea detrás de la parametrización de la señal, se refiere a que se requiere hacer una codificación de la señal comprimiendo los datos. La razón principal para realizar tal compresión, consiste en que en la medida que se trabaje con menos valores por tramo de la señal, mayor será la velocidad de respuesta del reconocedor. En la actualidad y por años se han usado diferentes algoritmos de parametrización (análisis LPC, análisis cepstral, bancos de filtros, etc.) [1][8][39][57][58], con muy buenos resultados, sin embargo, el grado de incertidumbre en un reconocedor siempre es muy alto debido a las diferentes fuentes de conocimiento que están involucradas.

En el reconocimiento, es indispensable que la respuesta sea rápida, casi en tiempo real, por lo que un "buen " reconocedor realizará la parametrización en paralelo, con otra serie de actividades. Este paralelismo de actividades, puede dar una idea de la complejidad inherente a los reconocedores de propósito general.

¿Qué hace el reconocedor con los vectores de propiedades de la señal acústica?. Asocia en una forma probabilística secuencias parciales de vectores de propiedades de la señal de entrada, con modelos matemáticos construidos en un proceso de "entrenamiento". Como esos modelos matemáticos son obtenidos a través de secuencias de propiedades extraídas de realizaciones de elementos de palabras habladas, entonces, están asociados en la memoria del reconocedor con símbolos del lenguaje del usuario. Esto significa, que el reconocedor averigua con qué probabilidad cada uno de los modelos que tiene en memoria, representan a una secuencia de vectores de propiedades, y a través de un mecanismo de selección de máximo, escoge aquél modelo que representa a la secuencia de vectores, con mayor probabilidad, y por lo tanto, va identificando paso a paso los elementos de las palabras que están presentes en la señal [6][26][42][56][57].

Como se puede intuir este proceso es complejo, que implica actividades como: dotar al reconocedor de la capacidad de decidir qué secuencia parcial de vectores corresponde a un elemento de la palabra, puesto que, la señal de entrada es para el reconocedor una secuencia de

vectores que hay que analizar por trozos, y donde hay que decidir si esos trozos se corresponden o no con símbolos de las palabras. También, hay que considerar que los sonidos que aparecen en una frase hablada de manera natural, no son eventos aislados o discretos, sino por el contrario, ocurren con solapamiento por efectos de la coarticulación [1][39], lo que hace muy difícil implementar un mecanismo que pueda determinar cuándo empieza y cuándo termina un fono o una palabra. También, hay que tomar en cuenta que la voz presenta una gran variabilidad, debido a que los sistemas de producción de voz de las personas no son iguales, e incluso las señales de la voz de una persona varía de un momento a otro, dependiendo, por ejemplo, del estado de ánimo y de otras situaciones ambientales. Hay elementos incluso sociales que pueden influir en el rendimiento de un sistema de reconocimiento, como es la región de procedencia del locutor, si está hablando en un idioma que no es el de su lugar de origen, si se trata de un hombre, una mujer o un niño, si hay ruido estable o intermitente en el ambiente, si hay ruido en el canal de recepción de la señal, etc. Todos esos elementos generan incertidumbre en el proceso de reconocimiento.

El módulo de descodificación acústica puede ser tan complejo como se desee, es decir, puede ir desde contener modelos de frases completas, donde toda la señal de entrada la relacione con modelos de frases, hasta contener modelos de fonos, donde tendría que relacionar trozos de la señal de entrada con diferentes modelos de fonos. Este último, es el tipo de reconocedores que existen actualmente.

En resumen, el módulo de descodificación acústica, genera una secuencia de elementos del lenguaje, que dependiendo del grado de exigencia que se le haga al reconocedor, constituyen la entrada para el descodificador lingüístico. En la figura 2.3, esa secuencia de elementos está constituida por palabras.

2.4.2. EL Subsistema de Descodificación Lingüístico

En la figura 2.5, se pueden apreciar dos tipos de subsistemas lingüísticos; en el subsistema **a**, se supone que éste recibe una secuencia de palabras por parte del subsistema acústico, mientras que en el subsistema **b**, se supone que recibe una secuencia de fonos. Se trata de dos

esquemas típicos de descodificación lingüística en el campo del reconocimiento y del procesamiento del lenguaje natural [1][22][39][42][56].

El subsistema de descodificación lingüístico, en la actualidad, tiene la tarea de recibir de parte del subsistema acústico secuencias de unidades del lenguaje, sean éstas, fonos, trifonos, sílabas, semisílabas, etc. [16][21][42][56], luego, debe determinar si esas secuencias de unidades constituyen una palabra, una orden ó una oración válida dentro del contexto de la aplicación del reconocedor. Bajo esta forma de operación, el subsistema de descodificación lingüístico constituye la fuente del conocimiento lingüístico de los reconocedores.

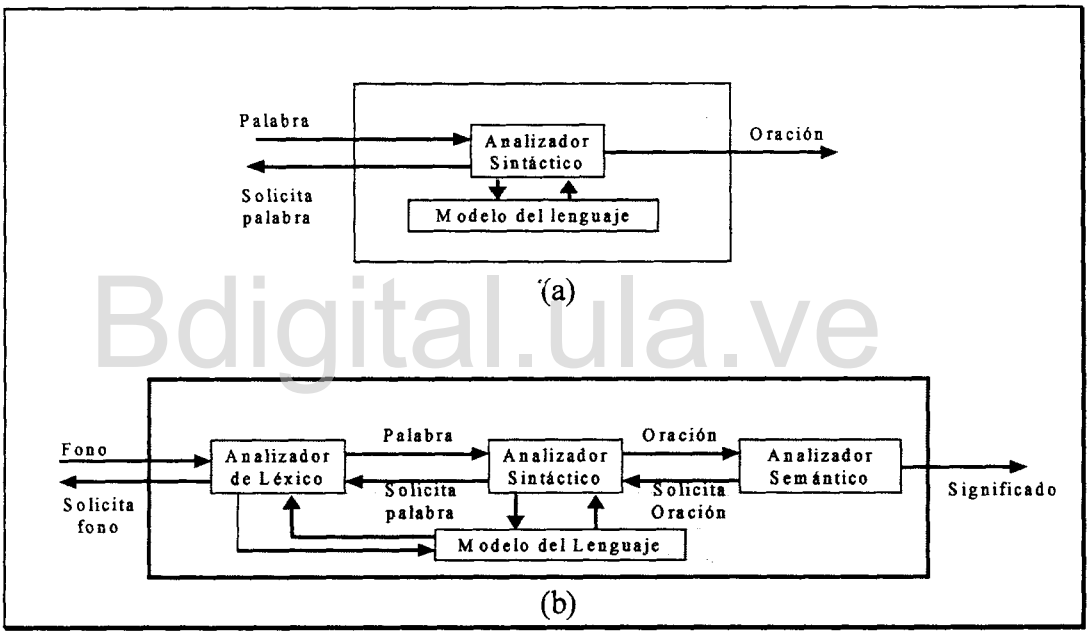


Figura 2.5. Sistemas de descodificación lingüístico, a y b.

Para poder llevar a cabo la función de detectar si la secuencia de unidades que recibe, de parte del descodificador acústico, son válidas, el descodificador lingüístico debe estar dotado de un modelo del lenguaje propio de la aplicación en la que se va a utilizar el reconocedor. El modelo del lenguaje se construye, siguiendo un enfoque parecido al utilizado para la construcción de los modelos acústicos, es decir, en una fase previa de entrenamiento.

La técnica más usada en la construcción de los modelos del lenguaje, consiste en seleccionar tantas oraciones, como sea posible, correspondientes al contexto o lenguaje de la aplicación para la cual se desea construir el reconocedor, luego, se recurre a un análisis estadístico de ese corpus, que consiste en obtener las frecuencias de ocurrencia de las palabras en el corpus, qué palabras pueden ser seguidas por otras y con qué probabilidad, qué palabras inician oraciones y con qué probabilidad, qué palabras finalizan oraciones y con qué probabilidad, etc., [1][31][39]. Con esta información, se construye un grafo, donde en cada uno de sus nodos se tenga la información de la presencia de una palabra y todas las palabras que las puedan anteceder y también, incluso partiendo de ese nodo se pueda obtener información sobre las palabras que la puedan suceder. Lo que se persigue, con un modelo de lenguaje en la forma de grafo, es construir una red de todas las posibles oraciones que puede recibir el reconocedor partiendo del conjunto de palabras distintas presentes en un corpus de oraciones y párrafos de entrenamiento.

Los modelos de lenguaje pueden estar constituidos por otros submódulos, como el que se describe a continuación tomando la idea de la figura 2.5.b. Puede existir un submódulo, que se encargue de detectar qué palabras serán formadas por las secuencias de fonos que produce el descodificador acústico, es decir, también puede haber un grafo adicional, cuyos nodos comprendan la representación ortográfica de los fonos, y una trayectoria completa dentro de éste constituiría una palabra. En ese caso, el descodificador tendrá una red de fonos que forman palabras conjuntamente con una red de palabras que forman oraciones.

Lo anterior significa que el modelo del lenguaje puede ser un componente de gran complejidad, y que existe una gran variedad de reconocedores dependiendo de la forma en que se implementen estos módulos [56].

Existen diversos algoritmos, bajo los cuales un descodificador lingüístico puede analizar si la secuencia de unidades recibida por parte del descodificador acústico, constituye una oración válida de la aplicación, sin embargo, el proceso general consiste en tomar la primera unidad, luego averiguar en el modelo del lenguaje, aquellas unidades que las pueden seguir con mayor probabilidad, es decir, se generan hipótesis de partes de palabras u oraciones que se puedan formar con las unidades de entrada. Luego, se analiza la siguiente unidad de entrada, en ese

momento se descartan algunas de las hipótesis previamente establecidas y se irán creando otras (el descarte se realiza, siguiendo el criterio de sólo tomar en cuenta, aquellas hipótesis que tengan una medida de probabilidad por encima de un umbral pre-establecido, en la medida que avanza el proceso de identificación). Este proceso, al inicio contendrá muchas hipótesis, pero en la medida que avanza el análisis de la secuencia se irán reduciendo, facilitando el proceso de identificación de las oraciones [22]. En paralelo con la construcción de la búsqueda de las hipótesis, el descodificador lingüístico lleva un registro del grado de confiabilidad del reconocimiento, que le permita indicarle al usuario, con qué nivel de confianza la secuencia que produce el descodificador acústico es una palabra o una oración válida para el lenguaje del reconocedor. Esto quiere decir, que si la secuencia de entrada, sigue un camino donde cada unidad sucede a la otra con una alta probabilidad, entonces al final se tendrá una secuencia con una probabilidad alta de ser una palabra o una oración válida del lenguaje de la aplicación.

También, se le pide en muchos casos al descodificador lingüístico, que cumpla la función correctora siguiente: supóngase que por ejemplo, producto de errores en el reconocimiento a nivel de fonos por parte del descodificador acústico, la secuencia forme una palabra equivocada, y por lo tanto, se genere una oración que no es válida para el lenguaje de la aplicación, se podría esperar que el descodificador lingüístico aprovechando su fuente de conocimiento, pueda sustituir la palabra que invalida la oración, por una que la transforme en válida. Para completar la explicación, supóngase que el descodificador acústico produce la siguiente secuencia de palabras de la señal acústica que recibe, "el Mérida me curó la herida", se podría esperar que el descodificador lingüístico fuese lo suficiente "inteligente" para indicar que la frase correcta corresponde a "el médico me curó la herida". Así, como en este ejemplo, se realiza una sustitución de una palabra, podría esperarse la inserción de una palabra que debido a alguna fuente de error no se halla detectado, o eliminar una palabra cuando debido a presencia de ruido, el descodificador acústico se confunda, e indique la ocurrencia de una palabra cuando en realidad no se produjo.

Hasta este momento podemos darnos cuenta, que el modelo del lenguaje impone condiciones al mensaje que le transmite el usuario al reconocedor, y que dependiendo de su complejidad admitirá mayor libertad en la construcción del mensaje. También hay que tomar en cuenta que en el proceso de reconocimiento aparecen diferentes fuentes de error, que van desde que se

produzca una mala pronunciación por parte del usuario, pasando por fallas en el subsistema acústico, en el modelado de las unidades acústicas, en el modelado del lenguaje y en los algoritmos de análisis lingüístico de las secuencias de unidades. Todo esto genera errores que hacen que los reconocedores actuales no sean tan robustos como se desea.

Para complementar lo descrito respecto al descodificador lingüístico, a continuación se presenta un resumen de la descripción que presenta Deller, et al [1], con respecto a una forma general de modelado del lenguaje para los reconocedores, como es el Modelo de Lenguaje de Peirce:

En el modelo de Peirce se supone que el descodificador acústico asocia los sonidos básicos del habla (fonos) con una serie de símbolos, eso sería la salida de este módulo. El descodificador lingüístico de Peirce está constituido por un módulo analizador de léxico a través del cual se forman las palabras, un analizador de sintaxis a través del cual se revisa la construcción de las oraciones (los dos constituyen el analizador de la gramática), un módulo analizador de la semántica y finalmente, un módulo analizador de la pragmática.

De la misma forma prevalece la idea de Reddy, según Deller et al [1], sobre el uso de las fuentes de conocimiento lingüístico en el reconocimiento de la voz: "Un locutor usa, sin darse cuenta su conocimiento del lenguaje, del ambiente, y del contexto para entender una oración.... las fuentes de conocimiento lingüísticas incluyen las características de los sonidos de la voz (conocimiento fonético), la variabilidad en la pronunciación (fonología), los patrones de entonación (elementos prosódicos), los patrones de sonidos de las palabras (elementos léxicales), la estructura gramatical del lenguaje (sintaxis), el significado de las palabras y de las oraciones (semántica) y el contexto de la conversación (pragmática)"

Implícitamente o explícitamente las fuentes de conocimiento lingüístico residentes en los reconocedores de voz, se pueden asociar con un componente del modelo de Peirce.

Para cerrar la discusión del funcionamiento de los descodificadores linguísticos, obsérvese la figura 2.5.b, donde se reciben de parte del descodificador acústico secuencias de representaciones de fonos. Existe un módulo que se encarga de determinar qué palabras estarán constituidas por las secuencias parciales de fonos (analizador de léxico), un analizador

sintáctico que se encarga de determinar qué oraciones del modelo del lenguaje forman las palabras obtenidas por parte del analizador de léxico; finalmente, un analizador semántico que se encarga de interpretar el significado de las oraciones.

Un ejemplo a través de cual se puede mostrar la función de un analizador semántico sería de la siguiente forma:

Supóngase que se quiere hacer una consulta a una base de datos a través de un sistema de reconocimiento de voz. Un usuario del sistema de información de la nómina de una empresa, podría desear la información relacionada con el sueldo que ganan los ingenieros que tienen entre 5 y 10 años en la empresa. Para obtener esta información a través de un sistema hipotético de reconocimiento para esa aplicación, el usuario podría generar los siguientes tipos de mensajes:

a.- Por favor muéstreme el sueldo de los ingenieros de entre cinco y diez años.
b.- Lísteme el sueldo de los ingenieros de cinco a diez años.
c.- El sueldo de los ingenieros de cinco a diez años por favor.
d.- Etc.

Un analizador semántico debería estar dotado de la "inteligencia", para determinar que lo que se quiere decir con esos mensajes, es que le pase al manejador de la base de datos la siguiente órden:

list ((empleados = ingenieros) **and** (duración>=5 **and** duración<=10)) sueldo.

Esta forma de la órden, donde aparecen las palabras del inglés "list" y "and", es la forma en la que trabajan la mayoría de los manejadores de bases de datos, por lo menos los que se usan en Venezuela [65][66].

2.5. VOCABULARIO DE LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICOS DE LA VOZ

Los sistemas de reconocimiento presentan como salidas palabras o secuencias de palabras, por lo tanto, manejan un vocabulario. Esto significa, que todo sistema de reconocimiento está dotado de un conjunto de palabras que componen su vocabulario, pero además, hay que tomar en cuenta que en este campo el término palabras, se refiere a cualquier tipo de pronunciación; y no necesariamente a la forma en que normalmente lo utilizan las personas. Es decir, en el contexto de los sistemas de reconocimiento, toda pronunciación diferente que pueda identificar el sistema se considera como una palabra diferente [39].

2.6. TIPOS DE RECONOCEDORES AUTOMÁTICOS DE LA VOZ

Los sistemas de reconocimiento se pueden clasificar de muchas maneras dependiendo de las aplicaciones para las cuales sean desarrollados. Hay sistemas de reconocimiento que según el número de palabras distintas que manejen, pueden ser considerados como de vocabularios pequeños, medianos, o grandes; otros, que según acepten como entrada la voz de una, dos, varias o muchas personas, se consideran como reconocedores dependientes o independientes del locutor; de la misma manera, otros que dependiendo de si su entrada contiene la pronunciación de una sola palabra, o varias palabras con pausas entre cada una de ellas, o sin pausas, sino pronunciadas de manera natural, serán reconocedores de palabras aisladas, reconocedores de palabras conectadas o reconocedores de habla continua [1][39], y finalmente para agregar otro tipo de estos sistemas, existen aquellos dotados de una gramática o sin gramática. Dentro de esta última categoría, se encuentran los que exigen que las secuencias de palabras que integran el mensaje, constituyan oraciones válidas del lenguaje o por el contrario, pueden constituir cualquier secuencia de palabras, sin importar el orden en que aparezcan.

2.6.1. Reconocedores de vocabularios pequeños, medianos y grandes

Los reconocedores se clasifican generalmente como de vocabularios pequeños, medianos y grandes. Sin embargo, no hay consenso en la literatura en cuanto al uso de esos términos, por lo que recogiendo la información que aparece en Déller, et al [1] diremos que, los reconocedores de vocabularios pequeños son aquellos que manejan entre 1 y 99 palabras; los reconocedores de vocabularios medianos son aquellos que manejan entre 100 y 999 palabras;

mientras que los reconocedores de grandes vocabularios son los que manejan más de 1000 palabras.

De todos modos, como existen reconocedores que manejan miles de palabras, 10000, 20000, 100000, 200000 (se requieren más de 200000 palabras para cubrir el idioma inglés, según Huang, et al [39]), entonces, no es extraño que un reconocedor de 1000 palabras pueda ser considerado en algunos contextos como de vocabulario pequeño. Lo que si se debe tener claro es que un vocabulario pequeño impone limitaciones a los usuarios, puesto que no le permite muchas libertades en cuanto a la escogencia de frases, y por lo tanto le resta flexibilidad al reconocedor. Pero por otro lado, un vocabulario pequeño tiende a dar lugar a mejores niveles de reconocimiento, debido a que en la toma de la decisión sobre qué palabras o frases le fueron suministradas, el grado de confusión es menor, ya que tiene menos salidas candidatas de donde escoger la correcta.

2.6.2. Reconocedores dependientes e independientes de los usuarios

Aquellos reconocedores que se entrenan con la voz de una sola persona, tienen un alto grado de reconocimiento para la voz de ese locutor, pero tienen la desventaja que si en algún momento se requiere que entiendan la voz de otro locutor es necesario, repetir el proceso de entrenamiento con la voz del nuevo locutor. A este tipo de reconocedores se les conoce como dependientes del locutor, mientras que, aquellos reconocedores cuyos modelos, son construidos con patrones de voz de un grupo considerable de personas, y que sean capaces de reconocer la voz de muchas personas, incluso de personas que no intervengan en la fase de entrenamiento, son los reconocedores independientes del locutor.

Existe un tipo intermedio de reconocedores, que pocas veces aparece en la literatura, que se puede llamar de múltiples locutores, que se entrena con la voz de un conjunto reducido de locutores, y es utilizado para reconocer la voz sólo de ese conjunto de personas. Este tipo de reconocedores son entrenados para manejar frases de un contexto particular, como por ejemplo, órdenes que se produzcan en una oficina y que las máquinas las pueden canalizar y por lo tanto, responder con ciertas actividades como búsqueda de información en alguna base

de datos. Desde nuestro punto de vista, a ese último tipo de reconocedores nos podemos referir también, como dependientes del locutor.

En general, los reconocedores dependientes del locutor presentan una capacidad de reconocimiento alta, comparados con los reconocedores independientes del locutor, pero su uso se orienta a aplicaciones muy específicas, mientras que los últimos pueden tener un uso más amplio. Con esto se quiere decir que, la construcción de uno u otro tipo de reconocedor depende de la aplicación, porque por ejemplo, un reconocedor dependiente del locutor puede facilitar el trabajo de una persona con alguna discapacidad en su voz, puesto que estará entrenado con la voz sólo de esa persona, mientras que un reconocedor que por ejemplo, atienda llamadas del público, necesariamente debe ser del tipo independiente del locutor.

2.6.3. Reconocedores de palabras aisladas, de palabras conectadas y habla continua

Los reconocedores cuyas señales a reconocer están constituidas por una sola palabra, se conocen como reconocedores de palabras aisladas. Estos reconocedores se caracterizan porque sus modelos de voz, modelan palabras y no otras unidades. Su uso es específico de las aplicaciones, donde se requiere como entrada la pronunciación de una palabra. Una aplicación de reconocimiento de palabras aisladas, podría ser una consulta a un diccionario electrónico, donde se desee recuperar el significado de una palabra particular.

También existen reconocedores cuyas señales a reconocer comprenden varias palabras, pero con una pausa suficientemente larga entre las pronunciaciones de cada una, que se conocen como reconocedores de palabras conectadas. En este tipo de reconocedores, los modelos de la voz también están constituidos por modelos de palabras como en el caso anterior, donde dichos modelos se construyen a partir de pronunciaciones discretas de las palabras, por lo tanto, no modelan los efectos alofónicos entre palabras, ni la articulación entre palabras [1][39]. En esta técnica de reconocimiento, una frase de entrada se descodifica poniendo en secuencia una serie de modelos y luego, se hace corresponder la frase con esos modelos concatenados. Para obtener buenos resultados se le solicita a los locutores que cooperen pronunciando las frases en forma lenta y cuidando la pronunciación de las palabras.

Una aplicación de los reconocedores de palabras conectadas, podría ser una consulta a una cuenta bancaria, donde se requiere alta precisión en el reconocimiento de los números de la cuenta, por lo tanto, la diferencia en tiempo entre las pronunciaciones de cada dígito debe ser de unos cuantos milisegundos, suficiente tiempo para detectar las pausas que permitan distinguir de forma clara la ocurrencia de una palabra, de la siguiente.

Existen reconocedores más complejos que los descritos, conocidos como reconocedores de habla continua, para los cuales los usuarios pronuncian el mensaje de una manera natural, o al menos con muy pocas restricciones. Estos reconocedores deben tener la capacidad de tratar los límites temporales de los sonidos, las diferentes fuentes de ruidos, anomalías en la señal de voz, los efectos coarticulatorios y el solapamiento de los sonidos en la señal acústica, por lo tanto deben ser suficientemente robustos como para detectar, cuándo en una frase el usuario se "come" un sonido o por el contrario "inserta" un sonido, y hacer la corrección respectiva. Algunos de esos casos se pueden apreciar como en el ejemplo siguiente de uno de los archivos de la SpeechDat Venezolana, donde el usuario pronuncia la frase: "veinticuatro e noviembre de mil novecientos cuarenticuatro". El reconocedor debería ser lo suficiente inteligente, como para darse cuenta que la frase que debe presentar como salida es: "veinticuatro de noviembre de mil novecientos cuarenta y cuatro". También, se puede tener el caso de una inserción por parte del usuario como en la siguiente frase, también de la Speechdat Venezolana: "en enero de mil novecientos ochenta y seis cinco". Aquí, el reconocedor debe ser capaz de borrar una palabra de la frase y mostrar que ésta debería ser, al menos de la siguiente forma: "en enero de mil novecientos ochenta y cinco".

Los reconocedores de habla contínua, por ser los más generales, son esenciales en aplicaciones donde numerosos usuarios tienen que interactuar con el reconocedor. Estos reconocedores son construidos en base a modelos de subunidades de palabras y no con modelos de palabras, lo cual se debe a que como se usan en aplicaciones donde los vocabularios son de tamaños grandes, la construcción de un modelo por palabra requeriría una gran cantidad de pronunciaciones de entrenamiento para el conjunto de los modelos, lo que haría que la tarea de modelado se convirtiera en un problema difícil de superar; por esta razón se recurre a la construcción de modelos de unidades como fonos, trifonos, semifonemas, sílabas, etc. [6][16][21][22][56], que constituyen conjuntos pequeños de elementos, como en el caso del

español peninsular, donde las unidades a entrenar a nivel de fonos o alófonos serían a lo más 31 [42]. Claro, que el uso de estas unidades implica el uso de algoritmos adicionales para formar las palabras y las oraciones, a partir de las unidades elementales modeladas, pero que tienen la ventaja de que se le puede incorporar a los reconocedores conocimiento fonológico, léxico, sintáctico e incluso relaciones estadísticas entre las unidades, y por lo tanto, se les puede incorporar alguna capacidad de interpretación de las frases.

Una aplicación de los sistemas de reconocimiento contínuo podría ser, el mismo ejemplo de la sección 2.4.2, donde se consulta a un sistema de información a través de pedirle el sueldo de los ingenieros de una empresa, por medio de una frase hablada. Otra aplicación, sería para redactar informes, donde se le dicte a la máquina hablando sin mayores restricciones en las pronunciaciones y donde se reciba la voz de cualquier persona.

2.6.4. Reconocedores dotados de gramática y sin gramática

Existen reconocedores que no contienen el módulo de descodificación lingüístico, por lo tanto no hacen ningún tipo de análisis lingüístico de las salidas que muestra el descodificador acústico, es decir, para estos reconocedores no es relevante averiguar si las unidades detectadas en la señal acústica constituyen palabras u oraciones válidas para un lenguaje dado. A este tipo de reconocedores se les conoce como reconocedores sin gramática.

Un ejemplo de un tipo de reconocedores sin gramática, sería el caso de una aplicación donde las señales acústicas estarían constituidas por secuencias de los dígitos, para consultar el saldo de tarjetas de crédito, donde los modelos acústicos serían modelos de palabras que representan a los dígitos. En ese caso, la respuesta del reconocedor podría contener cualquier combinación de la transcripción ortográfica de los dígitos.

Aquellos reconocedores dotados del descodificador lingüístico son los reconocedores con gramática, ya que éstos si realizan una revisión de las secuencias de las unidades de salida del descodificador acústico, con el fin de determinar si esas unidades siguen un orden y que por lo tanto constituyen palabras u oraciones válidas del lenguaje de la aplicación.

Un ejemplo de este tipo de reconocedores podría ser, un sistema que admita consultas cuyas entradas sean pronunciaciones de fechas, a través de las cuales se pueda obtener información sobre la edad de las personas de una determinada institución, si esas entradas son fechas de nacimiento. En este caso, supóngase que el reconocedor está construido en base a modelos de fonos, es decir, las salidas del descodificador acústico serían secuencias de transcripciones ortográficas de los fonos; el descodificador lingüístico estaría dotado, por un lado, de un módulo de análisis lexical para construir las palabras y por otro lado, de un módulo sintáctico que se encargaría de determinar si las palabras constituyen oraciones de fechas. Por lo tanto, el reconocedor contendría una gramática de fechas, a través de la cual averiguaría si la señal acústica se corresponde con pronunciaciones de fechas y podría dar además, una medida de esa correspondencia, e incluso, podría hacer correcciones del tipo inserción, modificación o borrado de palabras, para eliminar algunos errores que podrían haberse producido en el reconocimiento acústico, o hasta en la pronunciación realizada por el usuario.

2.7. CONSTRUCCIÓN DE LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

La construcción de los sistemas de reconocimiento comprenden varias etapas generales que se describen a continuación: como en toda ejecución de un proyecto, se comienza con la definición del alcance que se desea del reconocedor, luego hay que definir y diseñar los modelos de voz que se van a utilizar; en una etapa posterior hay que definir y obtener la base de datos de voz a través de la cual se realizará la construcción de los modelos y las pruebas iniciales de reconocimiento. Todo esto conlleva, por supuesto, la búsqueda y selección de las técnicas y algoritmos que, dependiendo de la dificultad de la tarea de reconocimiento, podrán ser usadas para el desarrollo de cada módulo del reconocedor y para la puesta en operación final.

El alcance se refiere al tipo de reconocedor que se desea construir, si se quiere un reconocedor de propósito general o por el contrario se tratará de un reconocedor que funcione dentro de una aplicación específica. En función del tipo de reconocedor que se desee, se procede a definir su vocabulario, y por lo tanto a construir la base de datos de voz. La construcción de la base de datos implica la recolección de las voces de los locutores, esta base de datos está determinada

por el alcance del reconocedor, es decir, depende de si se trata de un reconocedor independiente o dependiente del hablante, o si se trata de un reconocedor de habla contínua, de palabras aisladas o de palabras conectadas.

Conociendo el objetivo del proyecto y contando con la base de datos, hay otra etapa que cubrir; se trata de la construcción de los modelos de la voz, representativos de las unidades del lenguaje, que también están determinadas por la aplicación del reconocedor. Después de construidos los modelos de la voz, comienza la verdadera etapa de medición de la capacidad del reconocedor, que consiste en tomar pronunciaciones de test y presentárselas como entrada. Luego, se analiza una a una la respuesta que da el reconocedor a esas entradas desconocidas, es decir, se mide la capacidad de reconocimiento comparando la salida que presenta el reconocedor, con lo que realmente debería producir.

La construcción de los reconocedores se puede resumir en dos grandes fases conocidas como: fase de entrenamiento y fase de reconocimiento u operación.

2.7.1. Fase o etapa de entrenamiento de los sistemas de reconocimiento

El entrenamiento de los reconocedores comprende básicamente dos operaciones: la primera consiste en parametrizar todas las señales de voz que pertenecen a la base de datos, y la segunda, consiste en construir los modelos de la voz.

Para la parametrización de las señales, se recurre a diversas técnicas de procesamiento digital de señales, con el fin de buscar formas de comprimir cada señal y representarla a través de secuencias de vectores de parámetros, tratando en la medida de lo posible de minimizar la pérdida de la información en distintos dominios de análisis (temporal, frecuencial, cuefrencial, etc.).

Una vez que se dispone de ese espacio paramétrico de la base de datos de voz, se le divide en dos grupos de señales parametrizadas; un grupo que constituye el corpus de entrenamiento, que por lo general es de mayor tamaño, y otro grupo que constituye el corpus de test.

Luego, haciendo uso del corpus de entrenamiento, se recurre a diferentes técnicas (teoría de Modelos Ocultos de Markov, Redes Neurales Artificiales, Algoritmos Genéticos, Lógica Difusa, combinaciones de estos, otros, etc.) para crear modelos de unidades del habla. Cada uno de los modelos se construye en base a un conjunto de realizaciones propias del tipo de pronunciación que se quiere modelar, vale decir, si se va a construir el modelo de un determinado fono, entonces se utilizarán las pronunciaciones parametrizadas de ese fono, que estén presentes en el corpus de entrenamiento.

Este proceso de entrenamiento puede ser muy complejo y lento, dependiendo de las unidades del habla que se modelen, de la cantidad de datos disponibles y de las técnicas seleccionadas para la construcción de los modelos.

2.7.2. Fase de reconocimiento o etapa de prueba y puesta en operación

La etapa de prueba del reconocedor, consiste en presentarle a éste, señal por señal cada una de las secuencias de parámetros del corpus de test. Para cada una de esas señales, se observa su respuesta, que se refiere a la transformación de cada señal a una secuencia de grafemas. Luego, las secuencias de grafemas se comparan con las secuencias que deberían originarse (las referencias), si se alcanzara un reconocimiento del cien por cien.

Se acostumbra realizar el cálculo del nivel de reconocimiento, por secuencia, de las dos maneras siguientes [4]:

$$\text{Porcentaje de reconocimiento 1} = \frac{H}{N} * 100 \qquad (2.1)$$

$$\text{Porcentaje de reconocimiento 2} = \frac{H - I}{N} * 100 \qquad (2.2.)$$

Donde N representa el número total de unidades (por ejemplo, el número de palabras) presentes en la señal a reconocer.

H = N – D – S, es el número de unidades reconocidas correctamente.

D es la cantidad de unidades borradas.
S es la cantidad de unidades sustituidas.
I es la cantidad de unidades insertadas.

Los dos tipos de porcentajes de reconocimiento anteriores se obtienen primero por secuencia, y luego con todas las secuencias, lo que da el nivel de reconocimiento global del sistema.
Es importante, tomar en cuenta que las cantidades referidas como inserciones, borrados y sustituciones que se presentan en esta sección, son contadas al comparar los resultados de las salidas reales y las salidas de referencia, es decir, son errores del proceso de reconocimiento.

Los resultados de esta etapa de pruebas, puede dar lugar a la repetición de las fase de entrenamiento y a la preparación de nuevos datos, cuando sus resultados no convencen al diseñador. Una vez que se alcanzan porcentajes de reconocimiento suficientemente altos se puede dar por terminadas las pruebas, y se entra a la fase de operación final, que consiste en poner el reconocedor al servicio de la aplicación para la cual fue desarrollado.

Es bueno hacer notar que la mayoría de los reconocedores no logran porcentajes de reconocimiento muy cercanos al cien por ciento, a excepción de algunos del tipo de palabras aisladas y de pequeños vocabularios.

2.8. HERRAMIENTAS TEÓRICAS Y ALGORÍTMICAS UTILIZADAS EN LA IMPLEMENTACIÓN DE RECONOCEDORES

En la construcción de reconocedores automáticos del habla se recurre al uso de una serie de herramientas, que pertenecen algunas al mundo matemático clásico y otras al campo heurístico de la inteligencia artificial.

En ese conjunto de herramientas formales, dentro del campo de los reconocedores encontramos: los Modelos Ocultos de Markov, las redes Neurales Artificiales, diversos algoritmos de procesamiento digital de señales, algoritmos de clustering del tipo supervisado y

no supervisado, la lógica difusa, los algoritmos evolutivos, algoritmos sintácticos y estocásticos para el modelado del lenguaje y para el procesamiento del lenguaje natural, etc.

No es la intención de la tesis, explicar en qué consisten esas herramientas puesto que hay abundante literatura para ello. Sin embargo, en el anexo A se presenta una descripción detallada del área de los Modelos Ocultos de Markov por ser ésta la herramienta menos conocida en el País y porque tuvo un alto grado de uso en las pruebas que se desarrollaron en este trabajo. De la misma manera, en el anexo B se presenta una descripción general del análisis de Predicción Lineal y Cepstral debido a que constituyó la base de la parametrización que se realizó de las señales de voz.

CAPITULO III

LOS MODELOS OCULTOS DE MARKOV FRENTE A LAS REDES NEURONALES ARTIFICIALES: UN CASO DE ESTUDIO EN RECONOCIMIENTO AUTOMÁTICO DEL HABLA

3.1. INTRODUCCIÓN

En este capítulo se presentan los resultados de evaluar el rendimiento de las técnicas de los Modelos Ocultos de Markov (MOM) frente a los modelos de ANN Perceptrónicas Multicapas entrenadas con el algoritmo BKP [18][19][20][27][36][37], en un caso particular de reconocimiento automático de palabras aisladas. Las pruebas se realizaron con dos tipos de sistemas experimentales de reconocimiento: uno construido usando MOM y el otro construido usando ese tipo de ANN. Se comparan dichas técnicas en función de los resultados obtenidos en cada caso.

Los sistemas sobre los cuales se realizaron las pruebas, fueron desarrollados como parte de este trabajo y tienen un funcionamiento limitado, puesto que forman parte de una serie de programas que se están probando, y de los cuales se espera que en el futuro integren un sistema más general de reconocimiento que se pretende desarrollar en la Universidad de Los Andes.

Los resultados que se presentan en este capítulo, corresponden a la primera actividad práctica que se realizó como parte de la tesis en el campo del reconocimiento automático del habla. En general, estas pruebas perseguían dos objetivos: por un lado, probar el funcionamiento del software de reconocimiento desarrollado y por otro lado, observar el comportamiento de los MOM de Observaciones Discretas frente a un tipo de ANN.

3.2. BASE DE DATOS UTILIZADA

Se trabajó con una pequeña base de datos proporcionada por el Grupo de Procesado del Habla de la Universidad Politécnica de Cataluña, España. Dicha base de datos está compuesta por señales parametrizadas de pronunciaciones aisladas, de los dígitos catalanes de 10 personas [26].

El corpus de entrenamiento comprendía señales de 8 personas: 4 señales distintas por persona para cada dígito. Se disponía por lo tanto de 32 secuencias de parámetros por cada dígito, y de 320 secuencias para entrenar los diez dígitos.

El corpus de test estaba formado por señales de cada dígito, pronunciados por dos personas distintas a las personas cuyas pronunciaciones participaron en el corpus de entrenamiento. Cada una de esas dos personas realizó 8 pronunciaciones por dígito, es decir, se disponía de 16 secuencias de test por cada dígito y 160 para el test de los diez dígitos.

El vocabulario tratado fue entonces, el conjunto de los dígitos catalanes del cero al nueve. La razón por la cual se utilizaron esos datos, se debió a que para ese momento no se contaba con ninguna base de datos del habla venezolana para fines de reconocimiento, y porque para ese momento, era más importante probar los sistemas de reconocimiento que se habían programado, puesto que tampoco se contaba en la ULA con un sistema de desarrollo que permitiera hacer pruebas de reconocimiento.

3.3. FORMATO DE LA BASE DATOS DE LOS DIGITOS CATALANES

La base de datos de los dígitos catalanes estaba constituida por secuencias de parámetros indizados para cada pronunciación de los dígitos, es decir, las señales de voz no estaban en su forma digital original sino que ya se les había realizado el procesamiento que las dejaba prácticamente listas para construir MOM de observaciones discretas o cualquier otro tipo de modelos.

El procesamiento de dichas señales se resume a continuación: se les había realizado el proceso de parametrización, a través del cual se había generado por segmentos, una secuencia de vectores de propiedades. Los vectores de propiedades estaban constituidos por 12 parámetros cepstrales (Mel Cepstrum) más su primera y segunda derivada. Posteriormente, con el conjunto de todos los vectores de propiedades de todas las señales del corpus de entrenamiento, se construyeron 64 grupos de los vectores más parecidos en sentido de la métrica euclidiana por medio del algoritmo LBG de cuantificación vectorial [1][26][58][62]. Por lo tanto, se construyeron 64 vectores centroides, uno por cada grupo; a cada grupo se le identificaba con un símbolo ASCII. Debido a este proceso de cuantificación vectorial, cada señal de voz (pronunciación de un dígito), se representaba como una secuencia de símbolos del código ASCII.

El Grupo de Procesado del Habla de la Universidad Politécnica de Cataluña usó un cuantificador de 6 bits para codificar las pronunciaciones. Es decir, cada secuencia de vectores de propiedades se cuantificaba de forma que cada vector quedaba representado por un símbolo entre 64 posibles. Además, para codificar los 64 símbolos se escogieron los caracteres ASCII entre ";" y "z", ambos incluidos.

Esta es la forma en la que se recibió la base de datos de los dígitos catalanes, como un conjunto de archivos de texto. Esa forma resultó suficiente para por un lado, llegar a conocer las técnicas para crear Modelos Ocultos de Markov de Observaciones Discretas, y por otro lado, para hacer reconocimiento de palabras aisladas usando este tipo de modelos.

Por cada dígito habían dos archivos de texto, un archivo para entrenamiento y otro para test. Donde cada línea o registro de esos archivos era una secuencia de los símbolos asociados a los vectores de propiedades del dígito.

A manera de ejemplo, se presentan las cuatro secuencias de entrenamiento del dígito cero pronunciadas por el hablante uno:

MLMAAMMcsspsssscVVDD@CCCCJJRRa

MMMAAAAAMcppspppssccDD@@CCCCJJRRH;

]]MMAAAAMccspspsssscLccDD@CCCCCJRRRR

]BQQAMAAMMccpspspspscVVD@CCCCCCJJRRHH

3.4. ACONDICIONAMIENTO DE LOS ARCHIVOS DE LA BASE DE DATOS DE LOS DIGITOS CATALANES

Se convirtieron los archivos de texto (las secuencias de los símbolos) a archivos binarios de punto flotante símbolo a símbolo, para poder realizar los cálculos que se requerían para construir y probar los MOM, y las ANN perceptrónicas multicapas entrenadas con BKP.

3.5. CONSTRUCCIÓN DE LOS RECONOCEDORES BASADOS EN MOM

Al inicio del capítulo, se dijo que se trabajó con dos tipos de reconocedores. Efectivamente fue así, y además de cada tipo se construyeron por software varios reconocedores, como se apreciará más adelante.

Para el entrenamiento y las pruebas de algunos reconocedores basados en MOM, se utilizó el algoritmo Baum-Welch [1][6][8][13][14][17] y para otros, el algoritmo Viterbi [1][6][8][13][14] [17].

En lo sucesivo, para distinguir cuando se está hablando de un reconocedor construido y probado con Baum-Welch de uno construido y probado con Viterbi, se referirá a ellos como el sistema Baum-Welch y el sistema Viterbi.

Cada MOM correspondiente a un dígito se entrenó por separado, es decir, cada reconocedor basado en MOM creaba por separado los diez modelos de los dígitos.

3.6. CONSTRUCCIÓN DE LOS RECONOCEDORES BASADOS EN MODELOS NEURALES

El reconocimiento del corpus de entrenamiento por parte de los reconocedores que trabajaban con MOM de 5 a 8 estados fue del 100%, mientras que los reconocedores que trabajaban con menos estados por MOM, fallaban entre 1 y 16 secuencias (95% y 99.6875% de reconocimiento).

Respecto al reconocimiento del corpus de test, el mejor porcentaje se obtuvo cuando se señalaron 10 secuencias de forma errada (93.75% de reconocimiento) y el peor ocurrió cuando aparecieron identificaciones erradas de 23 secuencias (85.625% de reconocimiento).

En la tabla 3.2 se muestran los resultados de entrenar y probar 7 reconocedores Baum-Welch con topología del tipo Bakis [1][13]. Como resultado se observa que en los reconocedores que trabajan con MOM de 3 a 8 estados el reconocimiento fue del 100% para el corpus de entrenamiento, mientras que aquellos que trabajaban con 2 estados por MOM fallaban en 2 secuencias (99.375% de reconocimiento). Estos resultados son ligeramente mejores respecto a los obtenidos con los modelos ergódicos.

Con respecto al reconocimiento del corpus de test, el mejor porcentaje se dió cuando se señalaban 9 secuencias de forma errada (94.375% de reconocimiento) y el peor ocurrió cuando aparecieron identificaciones erradas de 14 secuencias (91.25% de reconocimiento).

Igual que en el caso de los modelos ergódicos, es importante notar que cuando se evalúan estos reconocedores con el corpus de test, el rendimiento no es 100% en ningún caso, sin embargo, los resultados aunque similares a los obtenidos con los modelos ergódicos son ligeramente mejores a esos.

En la tabla 3.3 se encuentran los resultados de utilizar MOM del tipo ergódico pero entrenados con el algoritmo Viterbi. Allí, se observa que los sistemas que trabajaban con MOM de 3 y 4 estados, reconocen 100% el corpus de entrenamiento, mientras que los sistemas que trabajaban con 2, 5, 6, 7 y 8 estados por MOM, fallaron entre 2 y 20 secuencias de las 320 (93.75% y 99.375% de reconocimiento). Respecto al reconocimiento del corpus de test, el mejor porcentaje de identificación se da cuando se señalan 14 secuencias de forma errada (91.25%

de reconocimiento) y el peor reconocimiento se da cuando aparecen identificaciones erradas de 34 secuencias (78.75% de reconocimiento).

Tabla 3.3. MOM ergódicos y Viterbi.

Estados	Corpus de entrenamiento	Corpus de test
2	93.75	83.125
3	100	90
4	100	91.25
5	99.375	87.5
6	97.8125	81.25
7	97.5	78.75
8	99.0625	83.75

Tabla 3.4. MOM Bakis y Viterbi.

Estados	Corpus de Entrenamiento	Corpus de Test
2	98.125	82.5
3	100	88.75
4	100	92.5
5	99.0625	87.5
6	100	82.5
7	98.75	85.5
8	99.375	83.125

Estos resultados indican que estos sistemas presentan una capacidad de acierto ligeramente inferior respecto a los resultados arrojados por los mismos modelos entrenados con Baum-Welch.

En la tabla 3.4 se muestran los resultados de entrenar y probar reconocedores con MOM tipo Bakis bajo el algoritmo de Viterbi. Allí, se observó que los sistemas que trabajaban con MOM de 3, 4 y 6 estados, reconocían 100% el corpus de entrenamiento, mientras que aquellos que trabajaban con 2 , 5, 7 y 8 estados por MOM fallaban entre 2 y 6 secuencias de las 320 (99.375% y 98.125% de aciertos). Respecto al reconocimiento del corpus de test, el mejor porcentaje se da cuando se señalan 12 secuencias de forma errada (92.5% de aciertos) y el peor ocurre cuando aparecen identificaciones erradas de 28 secuencias (82.5% de aciertos).

3.8. EVOLUCION DE LA VEROSIMILITUD CON LA RE-ESTIMACION DE LOS PARAMETROS EN LOS MOM

En la tabla 3.5 se presenta la evolución de la verosimilitud del MOM que se entrenó para representar el cero. Se trata de un modelo tipo Bakis de cinco estados para el cual se hicieron 12 re-estimaciones utilizando el algorítmo de Baum-Welch.

Tabla 3.5. Cálculo de la Verosimilitud con Baum-Welch

Iteración	1	2	3	4	5	6
Verosimilitud	-4301.14	-2822.11	- 2628.2	- 2577.75	-2503.32	-2479.09
Iteración	7	8	9	10	11	12
Verosimilitud	-2471.1	-2466.57	-2464.77	-2463.17	-2461.07	-2458.57

Allí, se puede apreciar que la verosimilitud de que dicho modelo represente las pronunciaciones del cero se hace mayor con cada iteración y da la impresión de que converge a un valor cercano a -2458. De hecho, es así, sin embargo la convergencia no es tan rápida como uno desearía, queremos decir con esto que podemos hacer veinte iteraciones, quizás más y la verosimilitud no llega a estar totalmente estable en ese número.

Para detener el entrenamiento se recurrió al criterio de que con una variación de la verosimilitud en un valor inferior al uno por ciento respecto a una iteración anterior se logran resultados aceptables [1][26][57]. En este caso particular, la variación ocurre de esa manera a partir de la sexta o séptima iteración.

Esta experiencia mostró que el comportamiento de la verosimilitud de los MOM usando Baum-Welch, es similar cuando se trabaja con modelos Bakis a cuando se trabaja con modelos Ergódicos.

En la tabla 3.6 se presenta la evolución de la verosimilitud del MOM que se entrenó para representar también al cero, con la diferencia respecto al caso anterior que se trata de un modelo ergódico de 5 estados para el cual se hicieron también 12 re-estimaciones por medio del algorítmo de Viterbi.

Allí, se puede apreciar como la verosimilitud de que dicho modelo represente las secuencias de entrenamiento del cero, se hace más pequeña con cada iteración y que converge de manera exacta al valor 3003.7. En esta oportunidad, la convergencia ocurre a partir de la cuarta iteración. De estos experimentos, también se observó que el comportamiento de la verosimilitud de los MOM usando Viterbi, es similar cuando se trabaja con modelos ergódicos

a cuando se trabaja con modelos Bakis.

Tabla 3.6. Cálculo de la Verosimilitud con Viterbi.

Iteración	1	2	3	4	5	6
Verosimilitud	5016.82	4188.72	3005.01	3004.92	3003.7	3003.7
Iteración	7	8	9	10	11	12
Verosimilitud	3003.7	3003.7	3003.7	3003.7	3003.7	3003.7

Algo, que no se puede dejar de mencionar es que para la implementación de los reconocedores Baum-Welch, se tuvo que realizar la normalización o escalado de las probabilidades, tal como se menciona en el anexo A, para evitar los problemas de precisión numérica que tienen esos algoritmos en su aplicación práctica.

3.9. PRUEBAS DEL RECONOCEDOR BASADO EN EL MODELO DE ANN PERCEPTRONICAS ENTRENADAS CON BKP

Los ensayos que se realizaron, así como los resultados obtenidos con tal modelo se resumen en las tablas 3.7, 3.8 , 3.9, 3.10, 3.11 y 3.12.

Tabla 3.7. Resultado de trabajar con la primera Red Neural.

Coeficiente de aprendizaje	**Valor de convergencia**	**Duración en pasos**	**Reconocimiento Ent. Y test**	
0.9	0.97	603	100%	41.875%
0.8	1.03	550	100%	43.75%
0.6	3.32	357	98.125%	37.50%
0.5	1.28	605	100%	40.625%
0.4	3.87	468	100%	36.25%
0.2	3.66	747	98.125%	41.25%
0.001	97.66	810	0%	0%

La tabla 3.7 muestra los resultados de entrenar la red neural perceptrónica multicapa cuyas

salidas deseadas, se muestran en la tabla 3.9. Específicamente, el ensayo consistió en entrenar la red con un coeficiente de aprendizaje dado y en observar cuando el error convergía a un valor; en el momento que se observaba la convergencia se detenía el entrenamiento, se contaba la cantidad de veces que era necesario presentarle el conjunto de entrenamiento para lograr tal convergencia, cuidando de evitar el sobre-entrenamiento; finalmente se evaluaba el porcentaje de aciertos al solicitarle identificar las secuencias de la base de datos. Se empleó esta forma de parada para evitar prolongar innecesariamente el entrenamiento, aunque para algunos ensayos se podría estar arriesgando un mejor rendimiento de la red, debido al comportamiento no lineal del error respecto al resto de parámetros que intervienen en la búsqueda del mejor modelo. Se podría presentar la situación en que después de estar en aparente convergencia, el error saltara a un mejor valor. Sin embargo, la mayoría de los ensayos mostraron que estos casos, las pocas veces que ocurrieron no daban mucha más información de la que se muestra en las tablas.

Para evitar el sobre-entrenamiento de la red, el error de convergencia se modificaba, y se repetía el entrenamiento.

Tabla 3.8. Resultados de trabajar con la segunda Red Neural.

Coeficiente de aprendizaje	**Valor de convergencia**	**Duración en pasos**	**Reconocimiento Ent. y test**	
0.9	11.25	445	93.75%	21.875%
0.8	11.33	306	93.75%	20.00%
0.6	11.76	425	94.07%	20.625%
0.5	12.96	415	94.07%	18.125%
0.4	11.88	479	92.81%	20.625%
0.2	9.95	747	93.75%	17.50%
0.001	116.38	810	0%	0%

Se puede observar en la tabla 3.7, que este modelo neural tiene una capacidad de reconocimiento cercana al 100% para el corpus de entrenamiento en la mayoría de los casos, mientras, que para el corpus de test es aproximadamente el 50%. También se puede observar

que con valores muy pequeños para el coeficiente de aprendizaje, la red tarda mucho tiempo en aprender y hasta se puede concluir que no aprende, cuando este coeficiente está muy cercano a cero.

En vista de los resultados obtenidos con este modelo neural, se pensó que tal vez modificando las salidas deseadas y/o la topología de la red, se podría lograr un mejor reconocimiento. Por esta razón, se realizaron ensayos con otras dos estructuras de neuronas.

En la tabla 3.8, se presentan los resultados obtenidos aplicando los ensayos a un segundo modelo neural, cuya diferencia con el primero estaba sólo en que las salidas deseadas son las que se muestran en la tabla 3.10. Se puede observar que esta nueva red tampoco presenta buena capacidad de reconocimiento para la base de datos completa.

Tabla 3.9. Salidas deseadas de la primera red neural.

El dígito	Salida Deseada
0	0000000000
1	0000000001
2	0000000010
3	0000000011
4	0000000100
5	0000000101
6	0000000110
7	0000000111
8	0000001000
9	0000001001

Tabla 3.10. Salidas deseadas de la segunda red neural.

El dígito	Salida Deseada
0	0000000001
1	0000000010
2	0000000100
3	0000001000
4	0000010000
5	0000100000
6	0001000000
7	0010000000
8	0100000000
9	1000000000

Las pruebas y los resultados que se obtuvieron con un tercer modelo con cuatro neuronas en la salida, se presentan en la tabla 3.11. En la tabla 3.12 se muestran las salidas esperadas de esta tercera red.

De los resultados obtenidos usando los tres modelos neurales se puede afirmar que la selección

de la topología de una red multicapa perceptrónica para este tipo de aplicaciones no es fácil y hay que recurrir al ensayo y al error. También, se puede afirmar que el tipo de modelos de redes que se usaron no da buenos resultados para estas aplicaciones y que se debe experimentar con otras configuraciones y/o trabajar con una base de datos más grande. Otra cosa que hay que señalar, es que aun cuando se realizaron pruebas en las que se le presentaba el conjunto de entrenamiento varios miles de veces a estos modelos (estos datos no aparecen en estas tablas), los resultados no eran muy superiores a los señalados anteriormente.

En todo caso la mejor configuración es la tercera red y la de peores resultados es la segunda.

Tabla 3.11. Tercera Red Neural.

Coeficiente de aprendizaje	**Valor de convergencia**	**Duración en pasos**	**Reconocimiento Ent. y test**	
0.9	3.81	385	100%	41.87%
0.8	2.80	379	100%	45.00%
0.6	2.73	456	100%	39.375%
0.5	5.2	412	97.5%	41.875%
0.4	2.53	619	100%	41.875%
0.2	3.31	1027	100%	42.50%
0.001	97.95	810	0%	0%

Tabla 3.12. Salidas deseadas de la tercera NN.

El dígito	0	1	2	3	4	5	6	7	8	9
Salida Deseada	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001

3.10. RESULTADOS Y CONCLUSIONES

Comparando los resultados obtenidos con los MOM y los obtenidos con el tipo de ANN descrito, se encuentra que los mismos favorecen ampliamente a los MOM.

La convergencia o parada del entrenamiento de los sistemas basados en MOM y de los

sistemas basados en ANN, requiere un consumo de tiempo similar, sólo por el hecho de que el nivel de cálculo en los MOM es mucho más complejo, pero en realidad se hacen muy pocas iteraciones o re-estimaciones de los parámetros con esos modelos. Mientras que si se pretende bajar el error de aprendizaje en el caso neural, el tiempo sería significativamente mayor (ésto no es necesario como se apreció en estas pruebas, pues no se gana mucho en cuanto a mejoras en el reconocimiento).

Los modelos neurales utilizados en este caso, exigen que la longitud de las secuencias a procesar sean iguales, es decir, exigen que todas las pronunciaciones tengan la misma duración, algo imposible en la realidad. Los MOM no presentan esta exigencia.

Entre los algorítmos usados para entrenar los sistemas basados en MOM, el que logra la convergencia en forma más rápida, por lejos, es el algoritmo Viterbi.

El algoritmo Baum-Welch produjo el sistema con mayor capacidad de reconocimiento.

Los MOM Bakis generaron una capacidad de reconocimiento mayor que los Ergódicos.

Las pruebas y los resultados obtenidos permiten asegurar que los reconocedores basados en MOM, desarrollados en la Universidad de Los Andes presentan buen funcionamiento, por lo tanto se tiene la confianza de haber aprendido esta técnica, lo que sirvió y seguirá sirviendo para atacar proyectos más exigentes dentro del campo del reconocimiento automático del habla en la ULA.

CAPITULO IV

BASE DE DATOS DE VOZ VENEZOLANA PARA RECONOCIMIENTO AUTOMÁTICO

4.1. INTRODUCCIÓN

En este capítulo se da información sobre los datos utilizados para llevar a cabo las distintas pruebas de modelado y reconocimiento automático del habla de los venezolanos.

Es preciso recordar en este momento, que en la construcción de todo sistema de reconocimiento, una etapa importante tiene que ver con la escogencia de los datos que permitan elaborar los modelos de la voz y del vocabulario asociado al lenguaje, sobre el cual se evaluará su rendimiento. Por esta razón, es indispensable seleccionar muy cuidadosamente las señales de las pronunciaciones tanto de entrenamiento como de prueba.

La descripción de la base de datos que se presenta en este capítulo, es producto de un resumen que se realizó del informe SALA SPANISH VENEZUELAN Database for the fixed Telephone network elaborado por Moreno A. y Mora E., [2], donde se explica el proceso de construcción de la SPEECHDAT Venezolana, y en el cual la Universidad de Los Andes participó activamente. Se recomienda recurrir a dicho informe si se quiere profundizar respecto a cómo fue la construcción y conocer más de las características de esa base de datos.

4.2. DATOS UTILIZADOS EN LAS PRUEBAS DE RECONOCIMIENTO DEL HABLA VENEZOLANA

Los datos que se utilizaron para la construcción de los modelos del habla venezolana y a través de los cuales, posteriormente se realizaron las pruebas de reconocimiento automático, pertenecen a la base de Datos de voz que se ha llamado SPEECHDAT Venezolana.

4.3. ORIGEN DE LA BASE DE DATOS SPEECHDAT VENEZOLANA

La SpeechDat Venezolana es una base de datos de voz propiedad de la Universidad Politécnica de Cataluña de España, que fue creada bajo el marco del proyecto SALA (SpeechDat across Latin América [32]) de dicha Universidad. El proyecto SALA, tiene por objetivo construir ocho bases de datos que coleccionen voces del español y del portugués hablado en América Latina, con la finalidad de poner dichos datos al servicio de desarrolladores de máquinas que tengan la propiedad de recibir y entregar información en forma hablada.

Las bases de datos del proyecto SALA fueron diseñadas y creadas siguiendo las especificaciones del formato de construcción de bases de datos de voz, SpeechDat [46], [47]. De ahí viene el nombre de SpeechDat Venezolana.

La construcción de la SpeechDat Venezolana fue dirigida y soportada económicamente por la Universidad Politécnica de Cataluña (UPC), España; de hecho el proceso de transcripción y el formateo de las voces se realizaron en esa Universidad, mientras que el diseño del corpus y la recolección de las voces fue ejecutada en la Universidad de los Andes (ULA) Mérida, Venezuela, en cuyo trabajo participamos activamente. Esta base de datos fue revisada y validada por la compañía Holandesa SPEX [51].

4.4. DESCRIPCION DE LA SPEECHDAT VENEZOLANA

Las bases de datos de voz creadas bajo el proyecto SALA, grabaron las voces a través de líneas telefónicas fijas (dos líneas analógicas).

En el caso de la SpeechDat Venezolana, esta base de datos comprende registros de voz de 1000 locutores o hablantes; de cada uno se grabaron 44 tipos de pronunciaciones, que dieron lugar a una base de datos de 44000 archivos de voz. La base de datos está distribuida en 5 CD-ROM bajo el esquema ISO 9660.

4.4.1 Formato de los archivos de voz de la SpeechDat venezolana

Las señales de voz fueron muestreadas a 8KHz y codificadas como secuencias de 8 bits a través del formato mu-law, sin ganancia automática de control (no comprimido, recomendación CCITT G.711). Cada tipo de pronunciación se encuentra almacenada en un archivo separado; cada archivo de voz tiene asociado un archivo ASCII con información relativa a los datos del locutor y al formato de almacenamiento.

4.4.2 Nomenclatura de los archivos de voz SpeechDat venezolana

Los nombres de los archivos siguen las convenciones ISO 9660 para nombrar archivos (8 caracteres para el nombre y 3 para la extensión). Se usa el formato siguiente:

```
DD NNNN CN. DC F
```

donde:

Tabla 4.1. Formato de los archivos que integran la SPEECHDAT venezolana

DD	Código de identificación de las bases de datos del proyecto SALA: A4= Grabada por teléfonos fijos; B4= Grabada por teléfonos celulares; C4= Base de datos para verificación del hablante.
NNNN	número que indica la sesión o llamada registrada (0000-9999)
CN	CN (A1-Z9) indica el tipo de pronunciación grabada y el número de pronunciación de ese tipo.
DC	Código del País, un código de dos letras; EV = Español de Venezuela
F	Código del tipo de archivo O= Archivo de texto con las etiquetas ortográficas, U= Archivo de voz

A continuación se presenta un ejemplo de cómo se nombra un archivo SALA: A40000D1.EVU, es el nombre de un archivo de voz (U) grabado por teléfono fijo (A4) a la primera persona (0000) que llamó al sistema de grabaciones, cuya voz corresponde a la pronunciación del primer tipo de fecha (D1) y el español corresponde al venezolano (EV).

4.4.3. Tipo de pronunciaciones grabadas

A continuación, en la tabla 4.2, se muestran los tipos de pronunciaciones que se grabaron de las llamadas de cada locutor.

Tabla 4.2. Pronunciaciones de la SPEECHDAT Venezolana

Código de la pronunciación	Número de pronunciaciones por tipo	Tipo de pronunciación
A	1-6	6 palabras claves ó de aplicación
B	1	1 secuencia de 10 dígitos aislados
C	1	1 número de hoja (6 dígitos)
C	2	1 número telefónico (9-11 dígitos)
C	3	1 número de tarjeta de crédito (14-16 dígitos)
C	4	1 número de 6 dígitos de un conjunto de 150
D	1	1 fecha espontánea, la fecha nacimiento
D	2	1 fecha en palabras
D	3	1 una fecha relativa, ej. ayer
E	1	1 frase donde intervienen palabras claves o palabras de aplicación
I	1	1 dígito aislado
L	1	1 apellido deletreado
L	2	1 nombre de ciudad deletreado
L	3	1 secuencia de letras
M	1	1 cantidad monetaria
N	1	1 número natural
O	1	1 apellido de un conjunto de 500
O	2	1 nombre de la ciudad donde nació o creció el locutor
O	3	1 nombre de ciudad entre los 500 más frecuentes
O	5	1 nombre de empresa entre los 500 más frecuentes

O	7	1 apellido de un conjunto de 150
Q	1	1 pregunta donde la respuesta predominante es la palabra si
Q	2	1 pregunta donde la respuesta predominante es la palabra no
S	1-9	9 oraciones ricas fonéticamente
S	0	1 oración adicional
T	1	1 hora del día (en forma espontánea, la hora de la llamada)
T	2	1 frase de hora en palabras
W	1-4	4 palabras ricas fonéticamente

4.4.4. Plataforma de grabación de la base de datos

Las principales características de la plataforma de grabación se muestran en la tabla 4.3:

Tabla 4.3. Características de la plataforma de grabación

Interface:	Analógica
Tarjeta:	Dialogic Proline 2V.
Computador:	PC con Windows 98
Software de la Tarjeta:	Dialogic System Software
Servidor de llamadas	UPC programa ADA-D (Software de aplicación escrito en C)
Líneas:	2

4.4.4.1. Condiciones de grabación

Las llamadas se agrupaban por regiones: CENTRAL, ZULIANA, LLANOS, SUD_ORIENTAL, ANDES

Ambiente: Casa u oficina, casilla telefónica, y lugar público.

En la tabla 4.4. se indica el número de llamadas que se obtuvieron en los distintos ambientes.

Tabla 4.4. Número de llamadas recibidas y porcentaje en función del ambiente de la llamada.

Ambiente	Llamadas recibidas	Porcentaje
Casa u oficina	802	80.2%
Casilla telefónica	170	17%
Lugar público	28	2.8%

Cada bloque de 100 locutores en la base de datos tiene una distribución similar de sexo, edad, región dialectal y ambiente.

4.4.4.2. Servidor de llamadas UPC ADA

Se utilizó el programa ADA-D desarrollado en la Universidad Politécnica de Cataluña para controlar la recepción de las llamadas y para permitir, las grabaciones de la voz de los locutores.

El programa ADA-D está diseñado para trabajar con tarjetas Dialogic y está basado en el software para manejo de voz de Dialogic para Windows 98 y Windows NT. Este software de grabación, incluye un detector de voz/silencio. El tiempo de grabación por señal de la SpeechDat Venezolana es de 5 segundos.

4.4.4.3. Personas que intervienen en la base de datos

Los hablantes que intervienen en la colección de voces, son principalmente estudiantes de varias universidades de Venezuela y sus familiares. Este método de recolección de datos tuvo acceso a una gran cantidad de personas de varias zonas dialectales, por lo que de esta manera se consiguió un balance en edad y en sexo.

Para obtener la información de voces se distribuyeron hojas a los locutores. El contenido de las hojas consistía de texto que se debía leer y preguntas que se debían responder; cada hoja contenía un código que identificaba al locutor a quien se le hacía la grabación, por lo tanto se

podía contar el número de llamadas que realizaba, y sólo se registraba una por locutor. También se entregaba a los locutores una hoja de instrucciones que los guiaban durante las grabaciones (ver anexo C).

Se trató de obtener voces de todo el territorio nacional por zonas dialectales, por esta razón, se dividió a Venezuela según la propuesta de Mora E. [49], en las regiones dialectales que aparecen en la tabla 4.5.: Central, Zuliana, Llanos, Sud-Oriental y Andina.

Tabla 4.5. Regiones dialectales, descripción y llamadas obtenidas en cada región.

REGION	DESCRIPCIÓN	GRABACIONES
CENTRAL	Distrito Federal, Miranda, Carabobo, Aragua, Lara, Yaracuy, Falcón	203
ZULIANA	Zulia	211
LLANOS	Portuguesa, Guárico, Cojedes, Apure, Barinas	180
SUD_ORIENTAL	Sucre, Nueva España, Monagas, Anzoátegui, Delta Amacuro, Bolívar y Amazonas	196
ANDES	Tachira Mérida, Trujillo	210
TOTAL		1000

Características de los locutores

En la tabla 4.6, se muestra la información relativa a la edad y al sexo de los locutores.

Tabla 4.6. Distribución de locutores agrupados por edad y sexo

Grupos de edades	Número de locutores			Porcentaje del total
	Masc.	Fem.	Total	
bajo 16	2	5	7	0.7
16-30	204	272	476	47.6
31-45	182	148	330	33
46-60	111	66	177	17.7
Sobre 60	5	5	10	1
Total	504	496	1000	100

4.4.4.4. La Transcripción de los sonidos de la voz

La transcripción tanto fonética como ortográfica de los archivos de voz, fue realizada por la Universidad Politécnica de Cataluña, España. La transcripción fonética fue realizada en forma automática, por medio de un software desarrollado en la UPC (SAGA: Spanish Automatic Graphemes to Allophones Transcriber), y a través de la notación fonémica SAMPA (Speech Assessment Methods Phonetic Alphabet) [52]. Sólo los nombres propios y los nombres de las compañías fueron chequeados manualmente.

Las transcripciones fonéticas o fonémicas de las palabras, realizadas por esa Universidad se basaron en el dialecto de Caracas y en el español Castellano, (para las pruebas que se efectuaron en la tesis, se realizó una transcripción fonética SAMPA propia, en base a los sonidos del habla que se encontraron en las distintas regiones de Venezuela).

Para ilustrar lo anterior, a continuación se muestran varias palabras bajo los dos tipos de transcripciones fonéticas que hizo la UPC, la de Caracas y la del español Castellano:

Alcanzar	a l k a N s ‘a r	a l k a n T ‘a r
Cigarrillo	s I G a rr ‘I jj o	T I G a rr ‘i L o
Microfichas	m I k r o f ‘i tS a h	m i k r o f ‘i tS a s

Es importante resaltar que a algunos eventos acústicos que aparecen en los archivos y que no corresponden a voz, también se les hizo la transcripción. Ese tipo de sonidos fueron agrupados en cuatro categorías, donde las dos primeras categorías se refieren a eventos originados por el locutor y las dos restantes originadas por otras fuentes.

Esas cuatro categorías son:
Pausas producto de dudas del locutor al estilo mm, ah, eeh, umm, etc.
Ruidos propios del locutor como toser, respiración, estornudos, etc.
Ruidos estacionarios de fondo como producto de un carro encendido, conversaciones cerca del locutor, canales ruidosos, etc.

Ruidos intermitentes como golpearse una puerta, el repique de un teléfono, un grito de un niño, etc.

4.4.4.5. Contenido de la base de datos

La base de datos colecciona habla espontánea y habla producto de la lectura. Para el caso de la grabación de habla espontánea se buscaba que se pudiera registrar el acento del locutor, a través de frases que indicaran las fechas de nacimiento, el lugar donde creció, el número de la hoja de texto que le correspondió, y un conjunto de oraciones diseñadas por expertos lingüistas para permitir obtener ese tipo de información.

Cada llamada registrada en la base de datos producía 44 archivos de voz, 43 de los cuales eran obligatorios dentro del proyecto SALA.

A continuación se dan detalles del tipo de información que se obtuvo por cada locutor:

Palabras de Aplicación (A1-6): en la tabla 4.7 se presenta el conjunto de palabras, que representan palabras claves para aplicaciones, donde se haga reconocimiento con el lenguaje español venezolano. El A1-6 que aparece entre paréntesis, indica que por cada locutor se grabaron 6 pronunciaciones del tipo palabras de aplicación. Esto es lo que se refleja en la tabla 4.2, como el código y el número de pronunciaciones por tipo.

Tabla 4.7. Lista de las palabras de la aplicación

1	Español	16	Siguiente
2	Finalizar	17	Final
3	Opciones	18	Insertar
4	Ayuda	19	Cambiar
5	Cancelar	20	Borrar
6	Parar	21	Guardar
7	Continuar	22	reproducir
8	Repetir	23	Grabar

9	Operadora	24	Enviar
10	Llamar	25	programa
11	Marcar	26	Número
12	Volver a marcar	27	asterisco
13	Directorio	28	transferir
14	Lista	29	desconectar
15	Anterior	30	Venezuela

Dígitos aislados (I1)

Cada locutor pronunció un dígito en forma aislada.

Cadena de Dígitos (B1)

Cada locutor leía una secuencia de dígitos dejando una breve pausa entre cada dígito. Esa secuencia se seleccionó aleatoriamente entre 150 secuencias diferentes.

Dígitos conectados (C1-4)

El locutor pronunció 4 cadenas de dígitos haciendo una lectura normal:

Leía el número de la hoja (4 dígitos) y otros dos dígitos más, **C1**.
Leía un número telefónico incluyendo el código del área con 2, 3 o 4 dígitos, más otros 5, 6 ó 7 dígitos, **C2**.
Leía un número de tarjetas de crédito, **C3**, con el siguiente formato:

xxxx xxxx xxxC xxxx (VISA)
xxxx xxxxxx xCxxx (American Express)

Se diseñaron 150 números diferentes de tarjetas de crédito, en los que se consideraron todas las combinaciones de tres dígitos, todas las combinaciones de silencio inicial más 2 dígitos y todas las combinaciones de 2 dígitos y silencio final.

Se leía un número de 6 dígitos, **C4**. Este número se seleccionaba aleatoriamente entre una lista elaborada de 150 cadenas de 6 dígitos cada una.

Fechas (D1-3)

Cada locutor pronunciaba tres fechas diferentes: una fecha espontánea, una fecha leída que fue elaborada en una forma convencional venezolana y una fecha general o relativa.

La fecha espontánea correspondía a la fecha de nacimiento, **D1**.

La fecha leída era del tipo "día de la semana, día-mes-año", **D2**, sólo se utilizaban dígitos para el año. Cada locutor pronunciaba una fecha diferente.

Ejemplo: Viernes, cuatro de noviembre de 1971.

Las fechas leídas comprendían desde los años 1920 al 2029 y los días del 1 al 31, (las palabras usadas para pronunciar las fechas se encuentran en la tabla 4.8).

Tabla 4.8. Lista de palabras contenidas en las fechas leídas.

Y	De, del	Febrero	Diciembre	noviembre
Lunes	Martes	Abril	marzo	
Miércoles	Jueves	Junio	mayo	
Viernes	Sábado	Agosto	julio	
Domingo	Enero	Octubre	septiembre	

Se pedía al locutor que dijera una fecha relativa, **D3**. Se trabajó con 88 fechas relativas. Las palabras usadas para formar estas fechas incluyen los días desde el 1 al 31, y las palabras contenidas en las tablas 4.8 y 4.9.

Tabla 4.9. Lista de palabras contenidas en las fechas relativas y generales

Al	Nochebuena
Anteayer	Nuevo(Año)
Antes	Pasado
Año	Pascua
Ayer	Primero
De	Próxima
Dentro	Próximo
Día	Que
Días	Ramos
El	Reyes
Hoy	Santa (Semana)
Ida	Santo (jueves, viernes)
La	Semana
Mañana	siguiente
Mes	Viene
Mil	Y
Navidad	

Ejemplos de fechas relativas:

el día de Reyes
dentro de quince días
Semana Santa

Frases de aplicaciones (E1)

Hay 150 frases de aplicaciones. Las frases contienen una o dos palabras claves; y se crearon específicamente para aplicaciones de reconocimiento automático. El número promedio de palabras por frase es 5.

En la tabla C.1 del anexo C, se muestran esas frases.

Pronunciación de nombres/palabras (L1-3)

Cada locutor pronunciaba tres tipos de textos: un nombre de ciudad, un apellido y una cadena de letras. En el español venezolano no existe deletreo espontáneo, es decir, no es normal en general, deletrear en el español a diferencia de otros lenguajes.

La tabla 4.10, muestra los símbolos de las letras de un diccionario español, el nombre usual de la letra, el nombre alternativo y la frecuencia de aparición (en porcentaje) para cada letra en la base de datos. Las letras que aparecen el menor número de veces son la X y la W (379 y 377 repeticiones). Esas letras no son comunes en el español venezolano, y sólo aparecen en apellidos o en algún nombre de ciudad.

Tabla 4.10. Nombres, nombres alternativos de letras del español y la frecuencia obtenida para cada letra en el conjunto de 1000 locutores.

Letra	Nombre	Nombre Alternativo	Frecuencia esperada (%)
A	A		12.60
B	Be		2.53
C	Ce		3.59
Ch	Che	ce hache e	2.03
D	De		3.17
E	E		6.40
F	Efe		1.80

G	Ge		2.57
H	Hache		1.79
I	I	i latina	4.95
J	Jota		2.03
K	Ka		1.69
L	Ele		3.85
Ll	Elle	doble ele, ele doble, ele ele	2.05
M	Eme		2.81
N	Ene		4.61
Ñ	Eñe		1.81
O	O		6.33
P	Pe		2.85
Q	Cu		1.99
R	Erre	ere	6.06
S	Ese		4.35
T	Te		3.75
U	U		3.96
V	Ve	ve pequeña, uve	2.59
W	doble ve	doble uve uve doble	1.70
X	Equis		1.71
Y	i griega	Ye	2.02
Z	Zeta		2.40

Apellido (L1)

Cada locutor pronunciaba un apellido de un conjunto de 500 apellidos comunes de Venezuela.

Nombre de ciudad (L2)

Cada locutor pronunciaba un nombre de ciudad de un conjunto de 500 nombres comunes de Venezuela y de América Latina.

Cadena de letras (L3)

Cada locutor deletreaba una cadena de letras diferentes a las que aparecían en el apellido y en el nombre de ciudad. La cadena de letras era diferente para cada locutor.

Cantidad monetaria (M1)

Cada locutor pronunciaba una cantidad diferente.

Ejemplos pronunciaciones de cantidades monetarias:
xxx bolívares
xxx,50 bolívares
xxx bolívares y 50 céntimos
xxx bolívares con 50 céntimos
x.xxx bolívares
xx.xxx bolívares

Número Natural (N1)

Cada locutor pronunciaba un número diferente, con longitudes que iban desde 2 a 6 dígitos. El formato para los números es el estándar del español. Los miles son indicados por (.).
xx
xxx
xx.xxx
xxx.xxx

Nombre propio ó apellido (O1)

Cada locutor pronunciaba un apellido o un nombre entre 500 comunes de Venezuela. Este dato es del mismo tipo que L1.

Nombre de ciudad en forma espontánea (O2)

El locutor decía la ciudad donde creció.

Nombre de ciudad (O3)

Cada locutor pronunciaba un nombre de ciudad de 500 de las más frecuentes en Venezuela y América Latina.

Nombre de empresas (O5)

Cada locutor pronunciaba un nombre de empresa de 500 posibles. El conjunto incluye nombres completos, agencias extranjeras, acrónimos, etc. No se daba indicación de cómo pronunciar.

Nombre y Apellido (O7)

El locutor pronunciaba un nombre y un apellido, entre 150 de los más comunes que hay en Venezuela.

Respuestas Si/no (Q1-2)

Se le hacía una pregunta al locutor cuya respuesta esperada era No.
Se le hacía una pregunta al locutor cuya respuesta esperada era Si.

Oraciones ricas fonéticamente (S1-9)

Se grabaron 5965 oraciones fonéticamente ricas diferentes. El número promedio de aparición de las oraciones es 1.5, y ninguna se grabó más de cuatro veces.

El corpus venezolano fue elaborado, a partir del corpus fonético español diseñado en los proyectos SpeechDat I y II de la Universidad Politécnica de Cataluña. Un grupo de lingüistas venezolanos de la Universidad de Los Andes en Mérida, Venezuela, modificaron esas oraciones para adaptarlas al español venezolano. Cambiaron el léxico y agregaron las palabras necesarias, para lograr un balance fonético de la base de datos con una versión del conjunto SAMPA, al que se puede llamar SAMPA venezolano.

La tabla 4.11 muestra la ocurrencia de los monofonos en las oraciones ricas fonéticamente del corpus. Se incluyen algunas variaciones alofónicas que pertenecen a los mismos fonemas: D-d, B-b, G-g, n-N.

Tabla 4.11. Ocurrencia de monofonos en el corpus de oraciones ricas fonéticamente.

Alófonos	Ocurrencias	Ocurrencias (%)
a	51490	12,87
e	49308	12,32
o	38116	9,53
r	23314	5,83
h	21946	5,49
N	21843	5,46
i	21639	5,41
l	20710	5,18
s	19285	4,82
t	17469	4,37
k	14256	3,56
d	11804	2,95
u	11730	2,93
p	10765	2,69
j	9304	2,33
D	9106	2,28
n	8981	2,24
m	8565	2,14

B	6539	1,63
b	3896	0,97
w	3854	0,96
f	3593	0,90
G	3266	0,82
rr	2984	0,75
jj	2536	0,63
g	1800	0,45
tS	1046	0,26
J	954	0,24
total	400099	100

Hora (T1-2)

Cada locutor pronunciaba dos horas diferentes:

Una hora espontánea, **T1**: se le pedía al locutor que indicara la hora al momento de la llamada.

Una hora pre-establecida, **T2**: Las palabras incluidas en las horas pre-establecidas se muestran en la tabla 4.12.

Tabla 4.12. Palabras incluidas en las frases de horas

A	exactamente	Mediodía	y
aproximadamente	Hoy	Minuto	un
Ayer	La	Minutos	dos
Cuarto	Las	Noche	cuatro
De	madrugada	Para	seis
Del	Mañana	(en) punto	ocho
En	Media	Son	diez
Es	medianoche	Tarde	doce
catorce	Cincuenta	diecisiete	treinta

dieciséis	Una	diecinueve	quince
dieciocho	Tres	Veintiún	uno
Veinte	Cinco	Veintitrés	veintiocho
Veintidós	Siete	veinticinco	trece
Veinticuatro	Nueve	veintisiete	cuarenta
Veintiséis	Once	veintinueve	

No se incluyó la palabra hora, debido a que sólo se usaba para preguntar, ni para responder. Tampoco se usó AM/PM, en su lugar se utilizó " de la mañana" o "de la tarde".

Ejemplos del formato para la hora pre-establecida:

exactamente a un cuarto para las dos de la madrugada de hoy
ayer a las dos y diez de la tarde

Palabras ricas fonéticamente (W1-4)

Las palabras ricas fonéticamente, fueron seleccionadas de un diccionario electrónico con 51000 raíces (Rodríguez S., [2]). El diccionario no contiene plural, variaciones femeninas de los adjetivos ni conjugaciones verbales. Primero se ordenaron las palabras y luego, se seleccionaron aquellas que contenían de 5 a 12 fonemas.

El conjunto de palabras fonéticamente ricas consta de 2400 palabras. La tabla 4.13, muestra el número de veces que aparecen los fonos en el conjunto de palabras ricas fonéticamente. El número total de fonos es 20783 y la tasa de repetición de cada palabra en la base de datos de los 1000 locutores es 1.6.

Tabla 4.13. Número de veces que aparecen

los fonos en el conjunto de palabras ricas fonéticamente.

fonema	Aparición
a	3182

e	2107
o	2048
r	1682
i	1264
t	898
N	894
h	865
s	743
k	722
l	694
jj	528
m	503
D	493
u	400
n	381
p	367
B	328
d	281
rr	274
b	271
j	270
tS	268
g	267
f	267
j	267
w	265
G	254

Una oración más (S0)

Para cada locutor se seleccionó una oración coloquial venezolana.

El léxico

El léxico comprende todas las palabras del corpus. Cada palabra distinta tiene una entrada separada y su codificación alfabética sigue las convenciones ISO-8859. La transcripción contiene la palabra, el número de veces que aparece en el corpus y su representación fonémica.

La tabla 4.14, incluye el conjunto completo de los símbolos SAMPA del español. El español hablado en Venezuela no usa los símbolos jj, dl, dZ, T, z, x, ts, C, S y Z.

Tabla 4.14. Conjunto de alófonos SAMPA Europeo y latinoamericano

SAMPA		Ejemplo	Transcripción
p	explosiva bilabial sorda	**p**ala	'pala
b	explosiva bilabial sonora	**b**ala	'bala
t	explosiva dental sorda	**t**ala	'tala
d	explosiva dental sonora	**d**ar	dar
k	explosiva velar sorda	**c**ala	'kala
g	explosiva velar sonora	**g**ala	'gala
m	nasal bilabial sonora	**m**ala	'mala
n	nasal alveolar sonora	**n**ada	'naDa
J	nasal palatal sonora	ca**ñ**a	'kaJa
tS	Africada palatal sorda	**ch**ico	'tSiko
f	fricativa labiodental sorda	**f**also	'falso
T	fricativa interdental sorda	**z**ona	'Tona
s	fricativa alveolar sorda	**s**ala	'sala
jj	fricativa palatal sonora	a**y**er	A'jjer
x	fricativa velar sorda	**j**amón	xa'mon
l	alveolar lateral sonora	**l**a	la
L	palatal lateral sonora	**ll**ana	'Lana
rr	alveolar vibrante sonora	ca**rr**o	'karro

j	palatal aproximante sonora	labio	'laBjo
w	labial-velar aproximante sonora	agua	'aGwa
B	Bilabial aproximante sonora	lava	'laBa
D	dental aproximante sonora	cada	'kaDa
G	velar aproximante sonora	lago	'laGo
r	alveolar vibrante simple	caro	'karo
a	vocal central abierta	tal	Tal
e	vocal anterior media	tela	'tela
i	vocal anterior cerrada	tila	'tila
o	vocal posterior media redondeada	todo	'toDo
u	vocal posterior cerrada redondeada	tul	Tul
N	velar nasal sonora	hongo	'oNgo
dl	lateral africada * sonora	náhuatl	‘nawadl
dZ	Palato alveolar africada sonora	cónyuge	‘kondZuxe
h	glottal fricativa sorda	pasta	‘pahta
ts	alveolar africada sorda	quetzal	ke'tsal
C	palatal fricativa sorda	cojín	Ko’Cin
S	palatoalveolar fricativa sorda	xocoyote	Soko’jjote
Z	palatoalveolar fricativa sonora	yugo	‘ZuGo

* para el español americano el símbolo dl implica africación

Los símbolos usados en la tabla 4.14 muestra las variaciones alofónicas. La siguiente lista muestra los fonemas correspondientes:

D y d pueden representar a /d/

B y b pueden representar a /b/

G y g pueden representar a /g/

jj y L pueden representar a /L/

n y N pueden representar /n/

4.5. RESULTADOS Y CONCLUSIONES

En la actualidad se cuenta con una base de datos del español hablado en Venezuela que sigue el formato Speechdat, coleccionada a través de líneas telefónicas fijas, que está al servicio de los investigadores del campo del reconocimiento automático del habla y de la lingüística, y muy particularmente al servicio de la Universidad de Los Andes. Es importante tener en cuenta, que esta base de datos es la única que existe actualmente, de voces venezolanas al servicio de la investigación en el campo de la Tecnología del habla y de la lingüística. También, es importante señalar que está en ejecución otro proyecto en el cual se está construyendo otra base de datos para reconocedores que reciban la voz venezolana través de teléfonos celulares. Dicho proyecto está bajo nuestra coordinación, aquí en Venezuela y está financiado por la empresa española Applied Technologies on Language and Speech, ATLAS [68].

CAPITULO V

RECONOCIMIENTO AUTOMÁTICO DE SECUENCIAS DE DIGITOS DEL HABLA VENEZOLANA: RESULTADOS DE PRUEBAS REALIZADAS CON LA VOZ DE MUJERES Y LA VOZ DE HOMBRES

5.1. INTRODUCCIÓN

El objetivo de este capítulo, es mostrar los resultados obtenidos de un conjunto de pruebas de reconocimiento de pronunciaciones de dígitos conectados. Se realizó reconocimiento del tipo independiente del hablante, puesto que se trabajó con la voz de un grupo considerable de mujeres y hombres de Venezuela.

Las actividades realizadas consistieron en: preparación de las señales de entrenamiento y de prueba, construcción de los modelos de voz, construcción del reconocedor y pruebas de reconocimiento. Todas estas actividades se realizaron a través de HTK, es decir, los sonidos se modelaron haciendo uso de la teoría de los MOM de observaciones contínuas.

Una vez construido el reconocedor, se le presentaron pronunciaciones de secuencias de dígitos y se observó su capacidad para detectar las dígitos presentes en cada pronunciación.

Con este conjunto de pruebas se crearon modelos de esas unidades tan importantes del habla, específicamente del habla venezolana, como son las pronunciaciones de los dígitos, de los cuales es fácil imaginar la gran aplicabilidad de un sistema que sea capaz de recibir información en forma de secuencias de dígitos pronunciados por cualquier venezolano.

5.2. JUSTIFICACIÓN DE LAS PRUEBAS

Las pruebas se realizaron con la idea de determinar si sería posible construir modelos de palabras suficientemente robustos, que permitieran reconocer automáticamente pronunciaciones de secuencias de dígitos, objetos útiles en aplicaciones donde la entrada a una máquina sólo requiere este tipo de pronunciaciones.

Los modelos que maneja el reconocedor comprenden entonces, características de pronunciaciones de los dígitos tanto de hombres como de mujeres. Se busca con este tipo de modelos, el desarrollo a futuro de algún reconocedor independiente del usuario, es decir, un reconocedor entrenado para trabajar sin importar que la voz provenga de cualquier persona que hable el español venezolano y donde la voz contenga sólo pronunciaciones de dígitos.

A nivel mundial, existe una gran variedad de trabajos relacionados con el tema [6][26][28][45], pero no se ha logrado todavía construir modelos de los dígitos y sistemas de reconocimiento de estos que sean cien por ciento confiables, cuando se espera que reciban la voz de cualquier persona, por lo que los distintos grupos de procesamiento han enfocado sus desarrollos a incorporar poco a poco modelos de los dígitos y de la voz en general, a sistemas que trabajen en ambientes controlados.

5.3. BASE DE DATOS DE SECUENCIAS DE DÍGITOS

Se trabajó en estas pruebas con 564 archivos de voz tomados de la base de datos SpeechDat venezolana, es decir, se trabajó con voz obtenida a través de líneas telefónicas fijas. Estos archivos contenían sólo pronunciaciones de secuencias de dígitos.

De los 564 archivos que se utilizaron en las pruebas, 426 fueron tomados como Corpus de Entrenamiento y 138 fueron tomados como Corpus de Reconocimiento. Las pronunciaciones de estos corpus fueron tomados de los siguientes estados y ciudades del País: Anzóategui, Sucre, Caracas, Carabobo, Miranda, Guárico, Yaracuy, Falcón, Zulia, Portuguesa, Mérida, Monágas, Barinas, Táchira, Trujillo, Bolivar, Aragua y Lara.

Las voces que se utilizaron en la etapa de reconocimiento, pertenecían a personas distintas a las que intervinieron en el entrenamiento.

5.4. UBICACIÓN Y PREPARACIÓN DE LOS ARCHIVOS DE VOZ PARA LAS PRUEBAS

La forma como se utilizaron los archivos de voz fue la siguiente: se localizaron dentro de la base de datos SpeechDat venezolana, un conjunto de pronunciaciones de dígitos de mujeres y hombres.

Posteriormente, cada archivo fue convertido del formato propio de la tarjeta de telefonía (DIALOGIC), con que se realizaron las grabaciones, a formato WAVE, con el fin de que el HTK [4] los pudiera procesar. Para esta conversión se utilizó el software VOXSTUDIO [52], una potente herramienta que permite la conversión entre diferentes formatos de archivos de audio.

Se realizó el etiquetado de cada uno de esos archivos de voz. Este etiquetado consistió en escuchar cada archivo de voz, y realizar la transcripción ortográfica de las pronunciaciones de los dígitos, de las zonas de silencio y de los ruidos, con el fin de identificar las realizaciones de cada una de las pronunciaciones que se deseaban modelar.

El proceso de etiquetado consistió entonces, en asociar los sonidos de los dígitos, detectados a través del oído, con secuencias de símbolos (palabras) del lenguaje español venezolano.

El último paso necesario para iniciar el modelado y la construcción del reconocedor, consistió en la parametrización de las señales presentes en cada archivo, para las que se obtuvieron los coeficientes MFCC (MEL Cepstrum [1],[48],[50]), la energía, la primera y segunda derivada, sobre segmentos de 25 milisegundos, enventanados a través de una ventana Hamming y desplazados cada 10 milisegundos sobre la señal de voz original; por lo que cada segmento se solapaba con el anterior en 15 milisegundos. Antes de la parametrización, cada señal se pasaba por un filtro de pre-énfasis cuyo coeficiente tenía el valor 0.98.

Para la construcción de estos modelos se emplearon vectores de 39 parámetros por segmento: 12 parámetros cepstrales, más la energía, con las respectivas primeras y segundas derivadas.

5.5. MODELOS DE LA VOZ UTILIZADOS

Los símbolos que se utilizaron para los modelos de los dígitos, estaban constituidos por su representación ortográfica: cero, uno, dos, tres, cuatro, cinco, seis, siete, ocho, nueve, sil y ru.

Como se puede apreciar, se trataba de modelos de palabras, donde aparte de los dígitos, se crearon modelos para las zonas de silencio (sil) y para los ruidos extraños de magnitud considerable (ru).

5.6. CONSTRUCCIÓN DEL RECONOCEDOR DE SECUENCIAS DE DÍGITOS

A partir de las 426 pronunciaciones de secuencias de dígitos del corpus de entrenamiento y de las realizaciones de cada palabra en dicho corpus, se crearon los modelos de los dígitos.

Aunque no se contaron cuántas realizaciones de cada dígito se utilizaron, un buen número estimado es alrededor de 400 realizaciones, ya que aunque en cada secuencia no aparecen siempre todos los dígitos, hay secuencias donde un dígito puede aparecer una, dos, o tres veces.

Los Modelos Ocultos de Markov que representan los dígitos son del tipo Bakis de 10 estados y se estimaron por medio del algoritmo Baum-Welch.

La escogencia del número de los estados, fue el resultado de un conjunto de pruebas de reconocimiento con pocas secuencias, donde se determinó que para los dígitos venezolanos, los MOM que daban mejores resultados eran los que tenían 10 estados.

En cuanto a la estructura tipo Bakis que se utilizó para los MOM, se debió a que ésta fue la estructura que mejores resultados arrojó en las pruebas que se efectuaron en el reconocimiento de los dígitos catalanes con MOM de observaciones discretas.

Inicialmente los modelos de los dígitos eran clones de un modelo prototipo que se creó con los vectores de medias y covarianzas globales, obtenidos a partir de todos los parámetros

extraídos de las señales del corpus de entrenamiento. Estos modelos inicialmente contenían una gausiana por estado.

Una vez que se obtuvieron esos modelos iniciales, el modelo de cada dígito se fue re-estimando progresivamente, con sus propias realizaciones hasta encontrar los mejores parámetros para cada modelo.

El proceso de búsqueda de los mejores parámetros de los modelos, consistió en hacer re-estimaciones y en posteriormente, observar la capacidad de reconocimiento con el corpus de prueba.

A continuación se describe el proceso de re-estimación de los modelos con los cuales se obtuvieron los mejores resultados:

A partir de los modelos iniciales descritos, se realizaron 3 re-estimaciones sucesivas a través de la versión "embedded training" del algoritmo de Baum-Welch [4]. Luego, se fue aumentando el número de gausianas por estado hasta llegar a seis gausianas. Cada vez que se aumentaba el número de gausianas, se efectuaban dos re-estimaciones del tipo indicado antes, a excepción del último caso donde con seis gausianas por estado, se realizaron 25 re-estimaciones hasta alcanzar un total de 36 por modelo. Este fue el número máximo de estimaciones que se realizaron, es decir, se hicieron pruebas con re-estimaciones que iban desde 3 a 36.

Para realizar las pruebas de reconocimiento, se programó el reconocedor a través de HTK.

5.7. DICCIONARIO Y GRAMÁTICA UTILIZADOS

En este caso el diccionario consistía sólo de una lista de palabras o representación ortográfica de los dígitos, es decir, el reconocedor estaba obligado a asociar cada representación acústica directamente con una palabra sin ningún paso intermedio. Por lo tanto, el modelo acústico que se detectaba era lo que el reconocedor mostraba directamente.

En cuanto al uso de gramática, realmente no se usó ningún tipo de gramática, puesto que se admitía que un dígito pudiera ir seguido por cualquier otro, como sucede en la realidad, que al escribir secuencias de dígitos, en cada posición puede ir cualquiera de éstos.

5.8. DESCRIPCIÓN DE LAS PRUEBAS REALIZADAS

Las pruebas que se realizaron con este reconocedor consistieron en presentarle como entrada, secuencias de pronunciaciones de dígitos y en averiguar su capacidad para identificar esas secuencias.

En las secciones 5.8.1 y 5.8.2, se describen las pruebas realizadas y se analizan los resultados obtenidos usando primero el corpus de entrenamiento y luego el corpus de test.

5.8.1. RECONOCIMIENTO DE PRONUNCIACIONES DE ENTRENAMIENTO

En esta sección se muestra los mejores resultados de las pruebas, en la que se le presentaron al reconocedor como entradas, las 426 secuencias del corpus de entrenamiento. Estos resultados fueron obtenidos con los modelos re-estimados 36 veces.

A continuación se presenta en forma resumida el resultado obtenido:

```
----------------------- Overall Results -------------------------
SENT: %Correct=75.35 [H=321, S=105, N=426]
WORD: %Corr=99.33, Acc=94.77 [H=2508, D=7, S=10, I=115, N=2525]
=============================================================
```

Como se puede apreciar, el reconocedor fue capaz de detectar en forma correcta 321 secuencias de las 426 que sirvieron para crear los modelos de las pronunciaciones de dígitos, y 2508 palabras de las 2525 que intervinieron en la prueba. Estos resultados produjeron un acierto del 75.35% para las secuencias completas y del 99.33% para las palabras presentes en dichas secuencias.

Para llevar a cabo estas pruebas, se comenzó trabajando realmente con más de 426 archivos, sin embargo, varios de éstos fueron descartados debido a que presentaban un alto nivel de ruido ambiental algunos, y otros contenían pronunciaciones de palabras distintas a los dígitos, como por ejemplo "cero uno siete éste ah perdón ocho".

En el anexo D se muestra, bajo el título **salidadigitosent4**, una parte mínima de la salida HTK que se obtuvo para esta prueba. No se muestra la salida completa debido a que el volúmen de páginas de la tesis se haría poco manejable; por esta razón, se mostrarán sólo unas pocas salidas de algunas pruebas donde se considera necesario presentar ejemplos (para la prueba que se acaba de describir, los resultados completos ocupan 19 páginas).

5.8.2. RECONOCIMIENTO CON PRONUNCIACIONES DE TEST

A continuación se presentan los resultados de las pruebas de reconocimiento que se realizaron con el corpus de test. En este caso, se muestra la capacidad del reconocimiento cuando los modelos fueron re-estimados 21, 27 y 36 veces con el corpus de entrenamiento.

5.8.2.1. Resultados obtenidos con los modelos re-estimados 21 veces

A continuación se muestran los resultados de hacer pruebas de reconocimiento con los 138 archivos de secuencias del corpus de test.

```
---------------------- Overall Results ---------------------------
SENT: %Correct=55.07 [H=76, S=62, N=138]
WORD: %Corr=98.56, Acc=88.28 [H=759, D=5, S=14, I=80, N=778]
================================================================
```

Se puede apreciar que el nivel de acierto de las secuencias nunca vistas es de 76 secuencias completamente reconocidas de las 138 que se le presentaron como entrada, para un porcentaje del 55.07% y un nivel de acierto de 759 palabras de las 778 que intervienen en la prueba, para un porcentaje de reconocimiento del 98.56%.

5.8.2.2. Resultados obtenidos con los modelos re-estimados 36 veces

A continuación se muestran los resultados de hacer pruebas de reconocimiento con los 138 archivos de secuencias del corpus de test, con los modelos re-estimados 36 veces.

```
---------------------- Overall Results -------------------------
SENT: %Correct=58.70 [H=81, S=57, N=138]
WORD: %Corr=98.43, Acc=88.95 [H=758, D=6, S=14, I=66, N=778]
================================================================
```

En este caso, se puede apreciar que al hacer un mayor número de re-estimaciones de los modelos, se consigue un mayor porcentaje de reconocimiento a nivel de secuencias completas (58.7%), que como se puede observar se logran reconocer completamente cinco secuencias más que en la prueba anterior, sin embargo, el porcentaje de reconocimiento a nivel de todas las palabras de la prueba, baja de 98.56% a 98.43%.

5.8.2.3. Resultados obtenidos con los modelos re-estimados 27 veces

A continuación se muestran los resultados de hacer pruebas de reconocimiento con los 138 archivos de secuencias del corpus de test, con los modelos re-estimados 27 veces.

```
---------------------- Overall Results -------------------------
SENT: %Correct=58.97 [H=80, S=58, N=138]
WORD: %Corr=98.56, Acc=88.17 [H=759, D=6, S=13, I=73, N=778]
================================================================
```

En este caso, el número de re-estimaciones de los modelos corresponde a una cantidad que se encuentra entre las cantidades utilizadas en las dos pruebas anteriores. Aquí, se consigue un porcentaje de reconocimiento a nivel de secuencias completas cercano al obtenido en la prueba con mayor número de re-estimaciones de los modelos, y un reconocimiento a nivel de todas las palabras de la prueba, igual al obtenido en la prueba con el menor número de re-estimaciones por modelo.

5.8.2.4. Resultados obtenidos con los modelos re-estimados 36 veces pero con un conjunto de 130 archivos de test

Se trata de una prueba donde se trabajó con el mismo corpus de test, con la excepción de que se eliminaron 8 archivos, debido a que al revisarlos se observó que presentaban un nivel de ruido alto.

Los resultados globales de esta prueba fueron:

```
---------------------- Overall Results ----------------------
SENT: %Correct=62.31 [H=81, S=49, N=130]
WORD: %Corr=98.57, Acc=91.89 [H=722, D=6, S=12, I=42, N=740]
=============================================================
```

Con ese corpus reducido, se logró como se esperaba un incremento considerable en el porcentaje de reconocimiento tanto a nivel de secuencias completas como a nivel de palabras.

Si se observa el porcentaje de reconocimiento tipo 1 (98.57%), permanece casi igual que en la prueba anterior, pero el porcentaje de reconocimiento tipo 2 (91.89%), se incrementó considerablemente, debido a que el reconocedor en este caso, cometió menos errores de inserción que en la prueba anterior. Los porcentajes de reconocimiento tipos 1 y 2 se explicaron en el capítulo 2.

5.8.2.5. Resultados obtenidos con los modelos re-estimados 36 veces pero con un conjunto de 126 archivos de test

Se trata de una prueba igual que la anterior, donde se eliminaron 4 archivos más, por su nivel de ruido alto.

Los resultados son los siguientes:

```
---------------------- Overall Results ----------------------
SENT: %Correct=64.29 [H=81, S=45, N=126]
WORD: %Corr=98.47, Acc=92.09 [H=710, D=1, S=10, I=46, N=721]
=============================================================
```

Como se puede apreciar, el reconocedor es capaz de detectar en forma correcta 81 secuencias de las 126 que se le presentaron y 710 palabras de las 721 que intervienen en la prueba. Estos resultados produjeron un acierto del 64.29% para las secuencias completas, del 98.47% y 92.09% (porcentaje tipo 2) para las palabras presentes en dichas secuencias, lo que constituye una mejora considerable respecto a las pruebas donde se trabajaba con los archivos ruidosos.

Parte de la salida HTK de esta prueba se encuentra en el anexo D, identificada como **salidadigitostest7**.

5.9. DISCUSIÓN DE RESULTADOS

Como es de suponer, en este tipo de experimentos se realizó una gran cantidad de pruebas, sin embargo, detallar cada uno de esas pruebas, es innecesario, por lo que los resultados que se muestran, son los mejores que se obtuvieron en esos experimentos.

Los resultados alcanzados son satisfactorios, ya que se logró obtener un alto porcentaje de reconocimiento a nivel de palabras, lo que se traduce en un reconocimiento promedio de secuencias completas también alto, lo que se puede observar en las salidas que se presentan en el anexo D, donde las fallas que aparecen en las secuencias que no son reconocidas completamente, se debe a que hay error en el reconocimiento de una o dos palabras de dicha secuencia.

De los resultados globales de las pruebas, se pudo observar que el nivel de reconocimiento obtenido con las pronunciaciones de un mayor número de personas, que el empleado en las pruebas descritas en [29] (15% para secuencias y 95.16% para palabras), es superior, lo que es una señal clara de que los modelos están mejor entrenados. También, el trabajo con transcripciones ortográficas alternativas generan mayor capacidad de reconocimiento a nivel de palabras, lo que indica que en general, el reconocimiento a nivel de las secuencias también es superior.

Por otro lado, se puede manifestar que los resultados obtenidos están en el rango cuantitativo obtenido por diversos grupos a nivel mundial en experimentos donde se hace reconocimiento de voz de palabras conectadas, como es este caso [26][45][50].

5.10. CONCLUSIONES

Se pueden crear modelos robustos de los dígitos pronunciados por venezolanos, sin embargo, para minimizar los errores de reconocimiento en las secuencias de éstos, en una aplicación real, habrá que recurrir a alguna técnica que minimice el porcentaje de error observado en estas pruebas. Una técnica que se nos ocurre es indicar al usuario que repita las secuencias un número de veces, y que el sistema determine cuál será la secuencia correcta.

Siempre habrá algún nivel de error de reconocimiento que hay que manejar, debido a la alta variabilidad de las señales de voz.

CAPITULO VI

RECONOCIMIENTO AUTOMÁTICO DE ORACIONES DEL HABLA VENEZOLANA: RESULTADOS DE PRUEBAS REALIZADAS CON FRASES PRONUNCIADAS POR MUJERES

6.1. INTRODUCCIÓN

En este capítulo se describe un conjunto de pruebas de reconocimiento automático de frases pronunciadas por personas de Venezuela; en particular, pronunciaciones de mujeres.

El procesamiento de las señales de la voz y el procedimiento de construcción de los modelos para realizar el reconocimiento, fueron en general, los mismos seguidos para realizar el modelado y el reconocimiento de las secuencias de los dígitos descrito en el capítulo anterior. La diferencia está en que los reconocedores que se desarrollaron en este caso, son reconocedores de habla contínua, independientes del hablante, donde se trabajó con la voz de mujeres; no de mujeres y hombres, aparte de que se crearon modelos de fonemas y algunas variaciones alofónicas de algunos fonemas, no de palabras.

Los reconocedores que se construyeron, son reconocedores de oraciones, y las oraciones que admiten son de tipo fechas, tal como se pronuncian y se utilizan en Venezuela.

Una vez construidos los reconocedores con HTK, se les presentaron pronunciaciones de fechas y se observó su capacidad para detectar los fonos, las palabras y la fecha presente en cada pronunciación. Para determinar sí las pronunciaciones eran fechas propias del venezolano, se dotaron a los reconocedores de una gramática que comprendía casi todas las formas en que los venezolanos pronuncian las fechas.

Se trató de un intento por construir modelos de la voz de mujeres de Venezuela, con el fin de adaptarlos a sistemas que en el futuro puedan realizar reconocimiento automático del habla de las mujeres de este País.

6.2. BASE DE DATOS UTILIZADA EN ESTAS PRUEBAS

Se trabajó con archivos de voz tomados de la base de datos SpeechDat Venezolana. Se seleccionaron archivos de pronunciaciones de fechas por parte de mujeres de 13 estados y ciudades de Venezuela. Se utilizaron 296 archivos, 214 fueron tomados como Corpus de Entrenamiento y 82 como Corpus de Reconocimiento.

Las pronunciaciones del corpus de entrenamiento, estaban distribuidas de la siguiente manera: 36 pronunciaciones de mujeres de Anzóategui, 13 de Sucre, 30 de Caracas, 30 de Falcón, 30 del Zulia, 15 de Portuguesa y 60 de Mérida.

Las pronunciaciones del corpus de test, estaban distribuidas de la siguiente manera: 4 pronunciaciones de mujeres de Barinas, 10 del Táchira, 8 de Trujillo, 6 de Bolívar, 10 de Aragua, 10 de Lara y 34 de Mérida.

6.3. UBICACIÓN Y PREPARACIÓN DE LOS ARCHIVOS DE VOZ

La forma como se utilizaron los archivos de voz fue la siguiente: se localizaron dentro de la base de datos SpeechDat Venezolana, un conjunto de archivos cuyos nombres siguen el formato A4XXXXDX.EVU, donde la A4 significa que el archivo se obtuvo por líneas telefónicas analógicas, la D refleja que la señal corresponde a una fecha, la X representa un dígito cualquiera, EVU es la extensión que indica que se trata de archivos del español hablado en Venezuela.

Posteriormente, para determinar sí la voz contenida en esos archivos correspondía a una mujer o a un hombre, se escuchaba cada archivo; de la misma manera se determinaba el lugar de origen de las personas, puesto que para cada persona a quien se le grabó la voz, hay un

archivo aparte (A4xxx xO2.EVU), con una frase donde se menciona el lugar donde ha transcurrido la mayor parte de su vida.

En paralelo con la identificación descrita en el párrafo anterior, se transcribía cada pronunciación de fecha en sus formas ortográfica y fonética, ésta última transcripción por medio, de una secuencia de símbolos fonéticos, los cuales aparecen en la sección 6.4. El procedimiento consistía en escuchar el archivo, luego se transcribían las palabras y posteriormente de ese texto resultante, se hacía su representación fonética en una versión SAMPA venezolana.

Posterior a ese tratamiento, cada archivo fue convertido del formato propio de la tarjeta de telefonía con la que se realizaron las grabaciones, a formato WAVE para el HTK.

El último paso necesario para iniciar el modelado y la construcción del reconocedor consistió en la parametrización de las señales, que se realizó siguiendo exactamente el mismo procedimiento utilizado para las pruebas descritas en el capítulo 5. La única diferencia respecto a esa parametrización, es que en este caso se trabajó, en algunas pruebas con 27 parámetros y en otras con 38.

6.4. MODELOS DE LA VOZ UTILIZADOS

Los modelos de voz que se construyeron para el reconocimiento de las pronunciaciones de fechas están al nivel de fonos.

En los archivos utilizados en las pruebas se encontró el siguiente conjunto de sonidos, al cual se ha llamado el Conjunto de Fonos de Fechas Venezolanas: a, b, B, c, d, D, e, f, g, G, h, i, j, k, l, m, n, N, M, o, p, r, R, s, t, u, w, y y sil.

Se puede apreciar que no aparece la v; ésto se justifica debido a que en la pronunciación venezolana no se distingue la ocurrencia de una v de una b. Lo mismo sucede con la s y la z. Tampoco aparece la x, puesto que tal fonema no está presente en el habla venezolana y menos aun, en pronunciaciones de fechas. El símbolo c se seleccionó para representar el sonido de la

ch y M para representar el sonido de la ñ. B, D y G representan la realización oclusiva (inicio de frase después de una nasal) de los fonemas /b/, /d/ y /g/, mientras que b, d y g representan la realización fricativa de los mismos fonemas. N representa la realización velar del fonema /n/ en distensión, es decir, al final de sílaba seguida por consonante. Se utilizó el símbolo sil para representar las zonas de silencio presentes en las pronunciaciones. El resto de símbolos representan sonidos bien definidos e identificados en los archivos.

6.5. ETIQUETADO DE LAS SEÑALES DE VOZ

En general, el etiquetado se realizó de la manera descrita en la sección 6.3, para el cual no era necesario alinear cada sonido con los símbolos, es decir, no se asociaba en forma manual y estrictamente cada símbolo con un segmento de la señal, ya que esto se dejaba para que HTK lo realizara automáticamente. A este proceso se le llamó segmentación automática.

Sin embargo, para los archivos de las fechas de las mujeres de Mérida se realizó ese tipo de etiquetado y también otra forma a la que se ha llamado segmentación semi-automática, que consistió en tomar cada archivo de voz, visualizar la señal en la pantalla, luego seleccionar segmentos de esa señal y escucharlos, identificar de qué fono se trataba, establecer los límites que lo separaban de sus vecinos inmediatos y hacer la transcripción simbólica respectiva. En este caso, se asociaba en forma directa cada fono con su respectivo símbolo, y también se establecía qué segmento de la señal completa correspondía, en cuanto a tiempo de duración, a cada sonido.

No se realizó el etiquetado en forma semi-automática de todas las señales que se utilizan en este tipo de experimento, debido a que esta actividad es demasiado lenta, tanto que por ejemplo, una persona relativamente experta etiquetaría unas 15 señales por día, trabajando 8 horas.

6.6. CONSTRUCCIÓN DE LOS MODELOS

A partir de las 214 pronunciaciones del corpus de entrenamiento y de las realizaciones de cada sonido en dicho corpus se crearon los modelos de los fonos.

El número de realizaciones utilizados por cada fono, para la prueba más general, se muestra en la tabla indicada como **Fonosmujeres**, que aparece en el anexo D.

Para este caso, los Modelos Ocultos de Markov que representan los fonos eran del tipo Bakis de 5 estados y se estimaron por medio del algoritmo Baum-Welch.

El número de estados por modelo, se escogió como producto de la orientación recibida por parte del Grupo de Procesado de la Voz, de la Universidad Politécnica de Cataluña, y por información que se encuentra en la literatura, donde aparece que un MOM adecuado para fonos, puede tener de 2 a 6 estados [1][6][21][57].

El proceso de construcción de los MOM de los fonos, es básicamente el mismo seguido para construir los modelos de los dígitos, con la diferencia de que en esta oportunidad se creó una cantidad mayor de modelos, aparte de que en algunos casos, hay muchos más datos para cada modelo, por lo que el proceso de re-estimación es más lento.

El proceso de búsqueda de los mejores parámetros de los modelos consistió en hacer re-estimaciones y en posteriormente observar la capacidad de reconocimiento con el corpus de prueba.

6.7. DICCIONARIO

Como se mostró en la sección 6.4, el sistema de reconocimiento está dotado de 29 modelos de sonidos del habla de mujeres venezolanas, lo que significa que dada una pronunciación como señal de entrada, el reconocedor debe ser capaz de identificar la secuencia de los sonidos presentes en esa pronunciación, es decir, hacer su transcripción fonética, y posteriormente, hacer la transcripción ortográfica de las palabras que están formadas por esos símbolos fonéticos.

Debido a que en este trabajo se planteó como objetivo hacer reconocimiento de palabras y oraciones, además del reconocimiento de fonos, se dotó al reconocedor de un diccionario cuyas entradas consistían de secuencias de sonidos (secuencia de fonos) y las salidas eran las

palabras ortográficas asociadas a esas secuencias de sonidos. En el anexo D se muestra, bajo el nombre **Diccionariofechas**, uno de los diccionarios utilizados en las pruebas, que contiene entradas alternativas.

El diccionario contiene entonces, la lista de las palabras que intervienen en las pronunciaciones de las fechas y los sonidos asociados a esas palabras.

En el diccionario se puede observar una misma palabra asociada con secuencias diferentes de sonidos, esto se debe a que se trató de cubrir para el reconocimiento, las diferentes variantes de pronunciación de las palabras que se detectaron en los archivos bajo estudio.

Para ilustrar esta situación, se presenta como ejemplo la palabra diciembre, que puede estar constituida por la secuencias de sonidos (fonos) siguientes: disjembre, Disjembre, disjemBre ó DisjemBre.

Lo anterior significa, que como las personas pronuncian las mismas palabras de manera diferente, entonces hay que buscar una forma adecuada, para asociar diferentes secuencias de sonidos (fonos) a una misma palabra. Esto se logra construyendo el diccionario en la manera descrita.

6.8. GRAMÁTICA

Así, como a través del diccionario se puede identificar la secuencia de palabras presentes en una pronunciación, para identificar oraciones hay que dotar a los reconocedores con las reglas gramaticales del lenguaje al cual están asociadas las pronunciaciones que se pretenda que éstos manejen. En este caso, se incorporó al reconocedor una gramática que consiste en la combinación adecuada de todas las palabras que usan los venezolanos para pronunciar o escribir fechas. En el Anexo D, aparece la programación de esa gramática bajo el nombre **grammujeres3**, la cual está escrita en el formato HTK.

Se debe tener presente, que el término gramática se usa en Tecnología del Habla, para señalar cualquier arreglo conveniente de símbolos; no se trata del sentido dado netamente en el campo de la Lingüística.

El programa que modela la gramática permite aceptar fechas de las formas que se muestran continuación, como ejemplo:

"El martes cuatro de diciembre de mil novecientos treinta y cuatro", "martes cuatro de diciembre de mil novecientos treinta y cuatro", "cuatro de diciembre de mil novecientos treinta y cuatro", "diciembre de mil novecientos treinta y cuatro", "cuatro doce de mil novecientos treinta y cuatro", "cuatro doce treinta y cuatro", "el cuatro doce de mil novecientos treinta y cuatro", "en el mes de diciembre de mil novecientos treinta y cuatro", "en diciembre del treinta y cuatro" y algunas otras.

6.9. DESCRIPCIÓN DE LAS PRUEBAS REALIZADAS

A continuación se describen las pruebas realizadas: en primer lugar, se describe la prueba de reconocimiento de fechas de las mujeres de Mérida, que constituyó el punto de arranque para la segunda prueba, que es más general, y que consistió en el reconocimiento de fechas de mujeres de Venezuela.

6.9.1. Pruebas con Voces de Mujeres de Mérida

Estas pruebas se realizaron a partir de reconocedores construidos en base a 60 archivos de entrenamiento y a 34 archivos del corpus de reconocimiento.

En este reconocimiento se realizaron dos tipos de pruebas: En el primer caso, la segmentación de la duración de los sonidos se hizo en forma semi-automática y en el segundo caso, dicha segmentación se realizó en forma automática. Para el primer tipo de pruebas, el mejor reconocedor fue construido en base a 15 re-estimaciones por modelo, donde se trabajó con 27 parámetros por tramo de la señal y para el segundo tipo de pruebas, el mejor reconocedor fue construido en base a 54 re-estimaciones por modelo y 38 parámetros por tramo.

6.9.2. Pruebas con Voces de Mujeres de Venezuela

Estas pruebas se realizaron a partir de reconocedores construidos en base a todos los archivos del corpus de entrenamiento y a 48 archivos del corpus de reconocimiento (no se utilizaron los 34 archivos de mujeres de Mérida de éste último corpus).

El etiquetado y la segmentación de los sonidos se hizo en forma automática. Se trabajó con 38 parámetros por tramo y el mejor reconocedor fue construido en base a 39 re-estimaciones por modelo.

En este caso, se realizaron pruebas separadas donde se utilizaron archivos que formaban parte del corpus de entrenamiento, y pruebas con archivos que formaban parte del corpus de reconocimiento. Los archivos para estas últimas pruebas fueron tomados de personas con orígenes distintos a los orígenes de las personas del entrenamiento.

Para este tipo de reconocimiento se trabajó en un caso, con la transcripción general o única para el habla Venezolana llamada Conjunto de Fonos de Fechas Venezolanas (ver sección 6.4), y en otro caso, con ese conjunto y algunas transcripciones alternativas adicionales, que son producto de las pronunciaciones que se han venido escuchando en la SpeechDat venezolana y de la experiencia diaria.

Todas las pruebas de reconocimiento se realizaron con el algorítmo Viterbi del HTK.

6.10. RESULTADOS

A continuación se describen los resultados más favorables obtenidos, en cada una de las pruebas.

6.10.1. Resultados de las pruebas de reconocimiento de fechas de las mujeres de Mérida

Para la primera prueba, la capacidad de reconocimiento resultó del 92.57% a nivel de palabras (324 palabras reconocidas correctamente de las 350 que intervienen en la prueba) y del 38.24% a nivel de fechas (13 fechas correctas de las 34 que intervienen).

En los resultados de la segunda prueba se encontró una capacidad de reconocimiento del 89.14% a nivel de palabras (312 palabras reconocidas correctamente de las 350 que intervienen en la prueba) y del 41.18% a nivel de fechas (14 fechas correctas de las 34 que intervienen en la prueba).

En estas pruebas se evaluó el reconocimiento en base a archivos de test solamente.

6.10.2. Resultados de las pruebas de reconocimiento de fechas de las mujeres de Venezuela

Para el caso donde se utilizaron transcripciones del Conjunto de Fonos de Fechas Venezolanas con 151 pronunciaciones tomadas del corpus de entrenamiento se encontró una capacidad de reconocimiento del 96.94% a nivel de palabras (1362 palabras reconocidas correctamente de las 1405 que intervienen en la prueba) y del 70.86% a nivel de fechas (107 fechas correctas de las 151 que intervienen en la prueba).

Para el caso donde se utilizaron transcripciones del Conjunto de Fonos de Fechas Venezolanas con 38 pronunciaciones tomadas del corpus de reconocimiento (después de revisar cuidadosamente las 48 pronunciaciones restantes sin tomar en cuenta las de Mérida, se encontraron archivos que contenían algunas palabras que normalmente no están presentes en las fechas y otros que tenían un nivel de ruido alto, por lo que fueron descartados 10 de esos archivos), se encontró una capacidad de reconocimiento del 95.88% a nivel de palabras (349 palabras reconocidas correctamente de las 364 que intervienen en la prueba) y del 63.16% a nivel de fechas (24 fechas correctas de las 38 que intervienen en la prueba).

Para el caso donde se utilizaron transcripciones alternativas y con las 151 pronunciaciones tomadas del corpus de entrenamiento se encontró una capacidad de reconocimiento del 97.01% a nivel de palabras (1363 palabras reconocidas correctamente de las 1405 que

intervienen en la prueba) y del 71.52% a nivel de fechas (108 fechas correctas de las 151 que intervienen en la prueba).

Para el caso, donde se utilizaron transcripciones alternativas y las 38 pronunciaciones de reconocimiento descritas, se encontró una capacidad de reconocimiento del 96.98% a nivel de palabras (353 palabras reconocidas correctamente de las 364 que intervienen en la prueba) y del 60.53% a nivel de fechas (23 fechas correctas de las 38 que intervienen en la prueba).

En el anexo D, se muestra parte del resultado de la última prueba identificada como **resultadosmujeres6**, y se resume los resultados de las restantes.

6.11. DISCUSIÓN DE LOS RESULTADOS

Los resultados alcanzados son satisfactorios, especialmente si se toma en cuenta que no sólo se estaba haciendo reconocimiento a nivel de fonos y de palabras, sino también de oraciones, y que para lograr este último tipo de reconocimiento, se requieren buenos resultados a nivel de las unidades previas, y eso sólo se logra cuando los modelos están bien entrenados y la gramática está bien definida. Otro aspecto a considerar a la hora de evaluar los resultados, es el hecho de que se trabajó con voces de muchos hablantes, lo que le da amplitud a los reconocedores, puesto que no es esclavo de la voz de una sola persona.

Para el caso del reconocimiento de fechas con voces de mujeres de Mérida, el resultado obtenido, cuando se utilizó el etiquetado y la segmentación semi-automática, es mejor en 3.43% a nivel de palabras, mientras que en el reconocimiento de frases es menor en un 2.94%. Esto último, podría suponer una contradicción, que no es tal debido a que así como hay frases en las que se disminuye la identificación correcta de algunas palabras, en otras puede aumentarse, y por lo tanto, algunas oraciones que antes resultaron no completamente reconocidas, es posible que posteriormente puedan ser reconocidas en su totalidad. Lo que pasa en esta prueba es que el nivel de reconocimiento de las palabras, en oraciones que en la segmentación automática resultaron no reconocidas completamente, aumentó pero no lo suficiente para que esas oraciones fueran completamente reconocidas. Por otro lado, en algunas oraciones que en la segmentación automática fueron completamente reconocidas, el

nivel de reconocimiento de las palabras disminuyó, lo que se traduce en que en la segmentación semi-automática, esas oraciones no se reconozcan completamente, por lo tanto, baja el nivel de reconocimiento de oraciones, aun cuando el porcentaje global de palabras reconocidas, aumenta.

Para el caso general, donde se hace reconocimiento de fechas con voces de mujeres de buena parte del territorio venezolano, se puede observar que la capacidad de reconocimiento, cuando se trabaja con algunas transcripciones alternativas para los sonidos, es 1.1% mejor a nivel de palabras al reconocimiento del caso cuando se trabaja con las transcripciones generales. Sin embargo, a nivel de frases completas, el nivel de reconocimiento con transcripciones generales, resulta mejor en 2.63%; este es el caso en que se trabajó con el corpus de reconocimiento, mientras que con el grupo de frases correspondientes al corpus de entrenamiento, el resultado estuvo siempre a favor del caso de transcripciones alternativas: 0.07% mejor a nivel de palabras y 0.66% mejor a nivel de frases. Desde luego, la contradicción aparente que se presenta cuando se trabaja con el corpus de reconocimiento, se explica como en el párrafo anterior.

De los resultados globales de las pruebas, se puede observar que el nivel de reconocimiento con las voces de mujeres de Venezuela es superior al obtenido cuando se trabajó con voces de las mujeres de Mérida, lo que es una señal clara de que los modelos están mejor entrenados.

También, el etiquetado y la segmentación semi-automática, así como el trabajo con transcripciones alternativas generan mayor capacidad de reconocimiento a nivel de palabras, lo que hace suponer que en general, el reconocimiento a nivel de frases también debe ser superior.

Por otro lado, se puede manifestar que los resultados obtenidos están en el rango cuantitativo obtenido por diversos grupos a nivel mundial, en experimentos donde se hace reconocimiento de voz contínua como es este caso [24][45][54][55][57][59].

6.12. CONCLUSIONES

Es posible crear modelos del habla de mujeres venezolanas a nivel de fonos para hacer reconocimiento automático de oraciones.

No se justifica el etiquetado y la segmentación semi-automática, si hay herramientas para hacerlo automáticamente lo que genera menor costo de trabajo, aun cuando los resultados que se obtengan puedan ser relativamente superiores.

Es posible crear reconocedores generales del habla de mujeres venezolanas, que puedan admitir la voz de muchas mujeres y hasta trabajar en forma independiente del hablante.

Los resultados obtenidos permiten suponer que en el futuro se podrán construir reconocedores del español hablado en Venezuela.

Se requiere una gran cantidad de pronunciaciones de entrenamiento por modelo, para obtener una capacidad de reconocimiento alta.

CAPITULO VII

RECONOCIMIENTO AUTOMATICO DE ORACIONES DEL HABLA VENEZOLANA: RESULTADOS DE PRUEBAS REALIZADAS CON FRASES PRONUNCIADAS POR MUJERES Y HOMBRES

7.1. INTRODUCCIÓN

En este capítulo se describen los resultados obtenidos de un conjunto de pruebas realizadas básicamente en las mismas condiciones de las pruebas descritas en el capítulo precedente, con la particularidad de que son más generales, en el sentido en que se estiman modelos de fonos que incluyen las características del habla, tanto de hombres como de mujeres.

Estas pruebas tienen el objetivo de generar modelos y por lo tanto, sistemas de reconocimiento que puedan recibir como entrada pronunciaciones de fechas de cualquier persona que hable el español venezolano, a diferencia del caso presentado en el capítulo anterior y el caso que presentaremos al final de este capítulo, donde la entrada debía ser exclusiva de mujeres en el primer caso y de hombres en el segundo caso.

7.2. JUSTIFICACIÓN DE LAS PRUEBAS

Como ya se ha mencionado en capítulos precedentes, la Tecnología del Habla viene realizando esfuerzos en distintos centros de investigación de muchos países con el objetivo de lograr una comunicación natural entre las máquinas y sus usuarios humanos. En ese sentido se han creado modelos de la voz de hombres, de mujeres y de niños; modelos de la voz pronunciada en distintos idiomas, modelos de la voz producida en distintos ambientes, modelos de la voz obtenida por distintos medios como micrófonos, teléfonos, etc, [51][54][61]. Esos esfuerzos orientados a la construcción de los modelos de la voz han producido resultados exitosos, y en la actualidad se sabe que existen sistemas comerciales que cumplen actividades útiles donde se

dan órdenes a las máquinas, se hace transcripción de texto, se hace consultas a algunos sistemas de información y hasta se establece algún tipo de dialogo con máquinas, todo por medio de la voz, donde se ha experimentado en la creación de diversos tipos de modelos como son: modelos de palabras, modelos de sílabas, modelos de fonos, modelos de trifonos, modelos de semifonemas, etc., [21][33][59].

Esa es la razón por la cual, en este trabajo se realizó un conjunto de pruebas en las que se crearon modelos que representan las características de un conjunto de fonos del habla de hombres y de mujeres, en modelos únicos.

7.3. BASE DE DATOS UTILIZADA EN LAS PRUEBAS

Para este conjunto de pruebas se trabajó con 332 archivos de pronunciaciones de fechas de hombres y mujeres de toda Venezuela (SPEECHDAT Venezolana). De los 332 archivos que se utilizaron en las pruebas, 260 fueron tomados como Corpus de Entrenamiento y 72 como Corpus de Reconocimiento.

Las pronunciaciones del Corpus de Entrenamiento estaban distribuidas de la siguiente manera: 140 pronunciaciones de mujeres y 120 de hombres.

Las pronunciaciones del Corpus de Reconocimiento estaban distribuidas de la siguiente manera: 38 pronunciaciones de mujeres y 34 pronunciaciones de hombres.

Las edades de las personas cuyas voces intervinieron en las pruebas oscilaban entre los 18 y los 60 años, con un grado de instrucción de educación media como mínimo.

7.4. PREPARACIÓN DE LOS ARCHIVOS DE VOZ

Todo el proceso de tratamiento de las señales de la voz, fue el mismo descrito en otras pruebas, es decir, se hizo la conversión a formato WAVE, luego las transcripciones ortográfica y fonética, por último el etiquetado y la segmentación, en este caso en la forma definida como segmentación automática.

La parametrización consistió en coeficientes cepstrales (MEL CEPSTRUM), igual que en los casos precedentes, y se trabajó con 38 parámetros por segmento.

7.5. MODELOS DEL HABLA VENEZOLANA UTILIZADOS EN LAS PRUEBAS

Los modelos del habla que se construyeron para el reconocimiento de las pronunciaciones de fechas estaban constituidos también por fonos, como se describió en el capítulo anterior.

El conjunto de fonos utilizado, fue el que se encontró en los archivos de entrenamiento y de reconocimiento, que correspondía exactamente al mismo descrito para el reconocimiento de fechas pronunciadas por mujeres, al que se ha llamado Conjunto de Fonos de Fechas Venezolanas. De esta manera se desea aclarar, que no se estaban tratando todos los fonos del habla venezolana, sino los que se utilizan en las fechas.

El número de realizaciones por fono utilizados en el entrenamiento, se encuentran en la tabla del anexo D identificada como **fonostotal**.

Los modelos se re-estimaron 39 veces siguiendo el mismo procedimiento descrito en el capítulo anterior.

7.6. PRUEBAS DE RECONOCIMIENTO DE LAS ORACIONES DE FECHAS

Aquí nuevamente, las pruebas de reconocimiento consistieron en presentarle, al sistema de reconocimiento, como entradas pronunciaciones de fechas y en observar la capacidad que tenía dicho sistema para identificar esas fechas.

Se realizaron en forma separada pruebas donde se utilizaron archivos que formaban parte del corpus de entrenamiento, y pruebas con archivos que formaban parte del corpus de reconocimiento. Los archivos para estas últimas pruebas fueron tomados de personas de lugares de origen distinto, al origen de las personas cuyas voces intervinieron en el entrenamiento.

7.7. EL DICCIONARIO Y LA GRAMATICA UTILIZADOS EN LAS PRUEBAS

La gramática es la misma utilizada en las pruebas precedentes donde se hizo reconocimiento de fechas, mientras que al diccionario se le agregaron algunas entradas alternativas.

7.8. RESULTADOS OBTENIDOS EN ESTAS PRUEBAS

A continuación se muestran los resultados más favorables obtenidos en esas pruebas.

7.8.1. Resultados de las pruebas de reconocimiento donde se trabajó con archivos del corpus de entrenamiento

Se realizaron varias pruebas con este corpus: inicialmente se hizo reconocimiento de los 260 archivos, en ese caso el sistema lograba reconocer completamente el 40% de las fechas y el 80.57% de las palabras que formaban esas fechas. Posterior a esta prueba inicial, se descartaron los archivos que contenían palabras que normalmente no se usan en las pronunciaciones de fechas, los archivos que presentaban altos niveles de ruido ambiental y se corrigieron aquellos que estaban acompañados de errores en las transcripciones.

Los resultados después de esa revisión, se muestran de manera resumida en la siguiente salida HTK.

---------- Overall Results ----------

SENT: %Correct=69.44 [H=150, S=66, N=216]

WORD: %Corr=96.94, Acc=94.96 [H=1963, D=15, S=47, I=40, N=2025]

==========

7.8.2. Resultados de las pruebas donde se trabajó con archivos del corpus de reconocimiento

El nivel de reconocimiento para los 72 archivos de ese corpus fue el siguiente:

---------- Overall Results ----------

SENT: %Correct=58.33 [H=42, S=30, N=72]

WORD: %Corr=93.69, Acc=88.77 [H=609, D=13, S=28, I=32, N=650]

==========

Se seleccionaron de los 72 archivos utilizados en la prueba anterior, 64 archivos que presentaban menores niveles de ruido ambiental, que contenían sólo palabras propias de las fechas y a algunos se les hizo algún tipo de corrección en la transcripción. A continuación se muestran los resultados de esa prueba:

---------- Overall Results ----------

SENT: %Correct=70.31 [H=45, S=19, N=64]

WORD: %Corr=96.82, Acc=94.14 [H=578, D=7, S=12, I=16, N=597]

==========

En el anexo D, se presenta parte de la salida de este último resultado, **en salidatotaltest3**.

7.9. DISCUSIÓN DE RESULTADOS

En esta prueba en que se hace reconocimiento a nivel de fonos, de palabras y de oraciones (fechas), en forma totalmente independiente del hablante, podemos darnos cuenta que los resultados son comparables a cuando se trabaja sólo con voces de mujeres, e incluso siendo rigurosos en la observación de los resultados, en este caso, el nivel de reconocimiento es menor, lo que lleva a suponer que los modelos se ven afectados por la variabilidad de la voz de los hombres, es decir, que aun cuando se ha incrementando el número de realizaciones por fono, los modelos parecen estar peor entrenados que en el caso cuando se trabaja con voces de mujeres.

De todos modos, estos resultados indican que es posible crear modelos del habla de los venezolanos a nivel de fonos, para hacer reconocimiento automático de oraciones y que en el futuro se podrán construir este tipo de reconocedores del español hablado en Venezuela, siempre que se trabaje con una gran cantidad de pronunciaciones de entrenamiento por modelo, para mejorar la capacidad de reconocimiento obtenida en estas pruebas.

7.10. RECONOCIMIENTO AUTOMÁTICO DE FECHAS PRONUNCIADAS POR HOMBRES

Para complementar las pruebas de reconocimiento de fechas, donde se trabajó en forma separada por un lado, con modelos de fonos de sólo mujeres y por otro lado, con modelos de fonos de mujeres y hombres, se realizaron un conjunto de pruebas adicionales, donde se hizo reconocimiento de fechas utilizando modelos de fonos creados sólo con la voz de hombres.

Las actividades realizadas consistieron exactamente en las mismas que se explicaron para los otros dos casos, sólo que las señales tratadas son de voz masculina.

Con estas pruebas se perseguía obtener modelos de la voz de hombres que pudieran ser incorporados a sistemas de reconocimiento de voz venezolana, que contengan en forma separada los modelos de la voz masculina y los modelos de la voz femenina, y que de esta manera se pueda contribuir a obtener una mayor tasa de reconocimiento si se estiman en los dos casos, modelos lo suficientemente robustos.

7.10.1. Base de datos utilizada en las pruebas de reconocimiento de fechas masculinas

Se utilizaron 138 archivos, de los cuales 99 fueron tomados como Corpus de Entrenamiento y 39 como Corpus de Reconocimiento.

Las pronunciaciones del Corpus de Entrenamiento estaban distribuidas de la siguiente manera: 9 pronunciaciones de hombres de Anzóategui, 15 de Caracas, 13 de Falcón, 13 del Zulia, 20 de Portuguesa, 14 de Táchira y 15 de Mérida.

Las pronunciaciones del Corpus de Reconocimiento estaban distribuidas de la siguiente manera: 2 pronunciaciones de hombres de Barinas, 10 del Táchira, 6 de Trujillo, 2 de Bolívar, 2 de Aragua, 4 de Lara, 5 de Portuguesa, 4 de Zulia y 4 de Miranda.

7.10.2. Modelos de la voz masculina utilizados

Los modelos de voz que se construyeron y al que se ha llamado Conjunto de Fonos de Fechas Masculinas Venezolanas fueron: M, N, R, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, w, y y sil.

Se puede apreciar que el conjunto de fonos distintos que se encontraron en las pronunciaciones, está compuesto por 26 sonidos; y que no aparecen las realizaciones oclusivas (B, D y G) de los fonemas /b/, /d/ y /g/ como en el caso del reconocimiento de las voces femeninas.

El hecho de que no aparezcan en el conjunto de fonos, las realizaciones D y G, no significa que los hombres no produzcan este tipo de sonidos, sino más bien se debe a que en estas pruebas se trabajó con corpus más pequeños. La justificación de uso de un corpus más reducido se debió a que se estaba haciendo uso de las primeras 500 llamadas registradas en la base de datos, y a que la identificación de los archivos se hizo muy pesada, debido a que había que escuchar archivo por archivo y revisarlos para descartar los que presentaban niveles altos de ruido, que no fueron pocos. Este problema también se presentó obviamente, en el caso de los archivos de mujeres, sólo que para esas pruebas, el trabajo fue más riguroso en cuanto al tiempo que se le dedicó.

Los mejores modelos se encontraron con 39 re-estimaciones.

7.10.3. Diccionario y gramática utilizados en el reconocimiento de fechas masculinas

La gramática utilizada es exactamente la misma de las pruebas precedentes, mientras que el diccionario presenta como única diferencia, que no hay palabras que presenten como entradas los símbolos que representan los sonidos de la D y de la G.

7.10.4. Los mejores resultados obtenidos en las pruebas de reconocimiento donde se trabajó con archivos del corpus de entrenamiento

Inicialmente se hizo reconocimiento de los 99 archivos, en ese caso el nivel de reconocimiento era muy pobre por lo que se fueron descartando para este reconocimiento, aquellos archivos

que contenían palabras no propias de las fechas y los archivos que presentaban altos niveles de ruido ambiental.

Finalmente, se realizó el reconocimiento con 65 archivos de entrenamiento con los que se obtuvo el siguiente resultado:

```
----------------------- Overall Results ------------------------
SENT: %Correct=13.85 [H=9, S=56, N=65]
WORD: %Corr=71.94, Acc=69.35 [H=446, D=64, S=110, I=16, N=620]
================================================================
```

Como se puede observar el nivel de reconocimiento es bastante bajo si se considera que se estaba haciendo reconocimiento con los archivos de entrenamiento.

7.10.5. Los mejores resultados obtenidos en las pruebas donde se trabajó con archivos del corpus de reconocimiento

A continuación, se muestran los resultados obtenidos cuando se trabajó con archivos de test:

```
----------------------- Overall Results ------------------------
SENT: %Correct=4.17 [H=1, S=23, N=24]
WORD: %Corr=50.68, Acc=46.12 [H=111, D=29, S=79, I=10, N=219]
================================================================
```

7.10.6. Discusión de resultados del reconocimiento de las fechas masculinas

Como se puede observar, los resultados en este caso son pobres debido a que la cantidad de patrones por fono es muy pequeño, lo que hace que esos modelos no sean estimados de manera adecuada. Esto se puede comprobar en **FonosMujeres** y **FonosHombres** en el anexo D, donde se observa que el número de realizaciones por fono, que aparece en el caso del modelado de fonos de mujeres es muy superior al caso del modelado de fonos de los hombres.

7.10.7. Conclusiones del reconocimiento de fechas masculinas

Para obtener buenos resultados de reconocimiento, se debe partir de modelos que estén lo suficientemente bien entrenados, y esto se logra recurriendo a un gran número de patrones de entrenamiento por fono. Del caso del reconocimiento de voz de mujeres venezolanas, se puede indicar que una cota inferior para el modelado de los fonos masculinos, podría ser la que aparece en **FonosMujeres** del anexo D.

CAPITULO VIII

RECONOCIMIENTO AUTOMÁTICO DE DIALECTOS VENEZOLANOS Y GENERO POR MEDIO DE MODELOS DE PALABRAS

8.1. INTRODUCCIÓN

En este capítulo se muestran los resultados obtenidos de un conjunto de pruebas en las que se crearon modelos de palabras, a través de los cuales se pretendía analizar la capacidad que podría tener un sistema de reconocimiento, para determinar si una pronunciación que se le suministraba como entrada, pertenecía a un hombre o a una mujer, y cuál era el dialecto de esa persona.

Se trató de un intento por hacer reconocimiento del género del locutor y de la zona dialectal venezolana a la que pertenecía ese locutor, entre cinco zonas modeladas.

8.2. MODELOS DE LOS DIALECTOS VENEZOLANOS Y DEL GENERO

Se esperaba que el sistema de reconocimiento diera suficientes indicios, que llevaran a tener la seguridad de que es posible construir en el futuro, reconocedores capaces de distinguir si la pronunciación que se le proporciona como entrada corresponde a un hombre o a una mujer de una de las cinco zonas dialectales siguientes: Zulia, Centro (Distrito Capital, Carabobo, Aragua, Yaracuy, Miranda, Lara, Falcón), Llanos (Portuguesa, Guárico, Cojedes, Apure, Barinas), Andes (Táchira, Mérida y Trujillo) y Sud-Oriente (Sucre, Anzoátegui, Monágas, Bolívar, Nueva Esparta, Amazonas y Delta Amacuro).

El criterio bajo el cual se pensó en esas cinco zonas dialectales fue en base al conocimiento que se tiene de la forma de hablar de las personas de los distintos estados y ciudades del País [2][49]. Esto por su puesto, está sujeto a error debido a que hay zonas o poblados colindantes

que se pudiese estar separando dialectalmente en unas de esas zonas y sin embargo, en la realidad pueden hablar de manera muy parecida.

Para la construcción de los modelos de los dialectos, se usaron secuencias de los dígitos pronunciados por personas naturales de las zonas mencionadas, es decir, para cada zona se crearon modelos de los dígitos.

Así, por ejemplo, para la zona andina se crearon los siguientes modelos: aceroh, aunoh, adosh, atresh, acuatroh, acincoh, aseish, asieteh, aochoh, anueveh, acerof, aunof, adosf, atresf, acuatrof, acincof, aseisf, asietef, aochof y anuevef.

Como se puede observar en el ejemplo, cada nombre que representa a un modelo de un dígito está precedido por una "a" que identifica a la región andina, y termina con una "h" o una "f", indicando que ese modelo corresponde a la pronunciación que realiza un hombre o una mujer. Esto significa que para la región andina se tienen veinte modelos de palabras.

De igual manera, para el resto de las zonas dialectales se tiene que los modelos de los dígitos para la región zuliana, comenzarán con una "z", los de la región central con una "c", los de la región llanera con una "ll" y los de la región Sud-oriental con una "o"; y todos terminaran con una "h" o una "f".

Por lo tanto, para este tipo de pruebas fue necesario construir 101 modelos, veinte por zona dialectal, más un modelo común para todas las zonas, que correspondía al modelo del silencio o pausa entre las pronunciaciones de los dígitos.

8.3. TOPOLOGÍA DE LOS MODELOS UTILIZADOS

Los modelos que se crearon para estas pruebas fueron MOM tipo izquierda-derecha, de 10 estados. El número de estados por modelo, se determinó como se indicó en el capítulo 5.

8.4. CONSTRUCCIÓN DE LOS MODELOS

Para la construcción de los modelos, se utilizó un conjunto de pronunciaciones de dígitos de hombres y de mujeres de cada zona dialectal. Así, para la región andina, se trabajó con pronunciaciones de secuencias de dígitos de 45 hombres y 45 mujeres, para la región zuliana 50 hombres y 50 mujeres, para la zona sudoriental 30 mujeres y 28 hombres, para la región central 49 hombres y 77 mujeres, y para la región llanera 17 hombres y 32 mujeres.

Hay que destacar también, que de la mayoría de esas personas, se utilizaron 3 pronunciaciones de secuencias de dígitos. Por lo tanto, para el caso de los 17 hombres de la región llanera, que es el caso donde se presentan menos realizaciones, en realidad se utilizaron cerca de 17*3 = 51 secuencias.

El total de archivos que se usaron en las pruebas alcanzó 707 secuencias de dígitos, de los cuales 138 se tomaron como archivos de test.

En cuanto a la parametrización de los archivos de voz, se obtuvieron de éstos los coeficientes cepstrales en escala MEL, la energía y las primeras y segundas derivadas, para un total de 27 parámetros por segmento.

La forma de la estimación de los modelos, partió de un modelo prototipo de medias y varianzas globales con los datos de la región andina. Luego ese modelo se entrenó con las realizaciones particulares de los dígitos de cada región y de acuerdo al género del hablante.

Cada MOM de los dígitos, inicialmente contenía una mezcla gausiana por estado, sin embargo, en la medida que se probaba la capacidad de reconocimiento se fue incrementando el número de mezclas por estado hasta un número de 12 mezclas. En cuanto al número de re-estimaciones de los modelos, se hicieron pruebas donde se alcanzaron hasta 50 re-estimaciones.

En cuanto al número de realizaciones por dígito, éste se puede apreciar en el anexo D, en **estadística10** (para el caso más general de reconocimiento de dialectos y género).

8.5. PRUEBAS REALIZADAS

Las pruebas de reconocimiento del dialecto y del género del hablante se realizaron de la siguiente manera:

8.5.1. Reconocimiento de dos dialectos

Se crearon los modelos de la región andina y los modelos de la región zuliana, y con éstos se realizaron pruebas donde se pretendía averiguar la capacidad del reconocedor para distinguir, cuándo la persona que le proporcionaba una señal de entrada era un hombre o una mujer de una de esas dos regiones.

Se escogieron estas dos regiones como punto de partida, debido a que el dialecto de estas personas es muy marcado y relativamente fácilmente distinguible por el oído humano venezolano. Se esperaba que la máquina lograra lo mismo.

8.5.2. Reconocimiento de cinco dialectos

Adicionalmente a los modelos andinos y zulianos, se crearon los modelos centrales, llaneros y sudorientales. En este caso, el número de archivos de entrenamiento fue de 426.

Este era un reconocedor bastante complejo, en el sentido de que contenía los 101 modelos descritos en la sección 8.2. Se esperaba que el nivel de reconocimiento bajara respecto a la prueba realizada con dos regiones, debido a que el grado de confusión se vería incrementado por el número de modelos por cada dígito.

8.5.3. Reconocimiento de tres dialectos

También se realizaron pruebas donde se utilizaron sesenta y un modelos, correspondientes a tres regiones dialectales.

8.6. RESULTADOS DE LAS PRUEBAS

El criterio que se utilizó para dar una medida de la capacidad del reconocimiento del dialecto y del género, al que pertenecían las secuencias de los dígitos, consistió en calcular manualmente dos porcentajes a partir de las salidas HTK: uno estaba determinado por el número de palabras que indicaban el género de la persona en la salida que presentaba el reconocedor y el otro, que indicaba de qué región era esa persona; llamémoslos porcentaje de género y porcentaje de dialecto.

Así por ejemplo, si la respuesta del reconocedor era la siguiente "aceroh zceroh asieteh acincoh aochoh ccuatrof", entonces se tenía que un porcentaje del 83,33% de las palabras indicaba que la secuencia era pronunciada por un hombre (el porcentaje de género), y un porcentaje del 66,66% indicaba que ese hombre era andino (el porcentaje de dialecto).

A continuación se presentan los mejores resultados obtenidos en las pruebas realizadas.

8.6.1. Resultados de las pruebas de los dos dialectos

En la siguiente salida que presenta el HTK, se muestran los resultados obtenidos de tomar al azar 22 pronunciaciones de las 211 que se usaron en el entrenamiento.

```
---------------------- Overall Results --------------------------
SENT: %Correct=54.55 [H=12, S=10, N=22]
WORD: %Corr=94.53, Acc=90.62 [H=121, D=2, S=5, I=5, N=128]
=============================================================
```

Una salida HTK para esta prueba tiene la forma siguiente:

Salida esperada: aceroh adosh acuatroh acincoh anueveh aochoh asieteh
Salida obtenida: aceroh zdosh acuatrof acincoh znueveh aochoh asieteh

Bajo el criterio explicado antes, y si se fija un porcentaje de acierto, en las palabras reconocidas por HTK en cada secuencia, por encima del 50% para tomar una decisión a favor de un dialecto y del género, entonces, es claro que la respuesta sería que la secuencia de

entrada fue pronunciada por un hombre de los andes, puesto que el porcentaje de género indica que el 85,714% de las palabras corresponden a un hombre, y el porcentaje de dialectos indica que el 71.43% de las palabras corresponden a una persona andina.

En el sentido explicado antes, en las pruebas de reconocimiento donde se trabajó con las regiones andina y zuliana, los resultados obtenidos fueron 100% reconocimiento de género y 100% de reconocimiento de dialecto cuando se utilizaron archivos de entrenamiento.

Con respecto a los resultados obtenidos cuando se utilizaron archivos de test, los mejores resultados se muestran en la siguiente salida HTK:

```
------------------------ Overall Results --------------------------
SENT: %Correct=7.41 [H=2, S=25, N=27]
WORD: %Corr=59.01, Acc=47.83 [H=95, D=7, S=59, I=18, N=161]
==================================================================
```


En este caso, el porcentaje global de reconocimiento a nivel de palabras que presenta HTK (el porcentaje 1, según se explicó en el capítulo 2) es 59.01% (superior al umbral mínimo fijado en 50%), luego, al calcular los porcentajes (de género y de dialecto) por cada salida, el reconocimiento a nivel del género alcanzó el 93.103%, mientras que a nivel de dialecto fue del 72.41%.

Para aclarar porqué se exige el umbral de 50% de reconocimiento de las palabras, por parte de HTK, para cada secuencia, supóngase que se obtiene una salida de la forma **aceroh zdosh zcuatrof zcincoh znueveh aochoh asieteh.** Supóngase también, que el reconocedor identifica correctamente sólo el 40% de las palabras presentes en esa secuencia, entonces la decisión de decir que esa secuencia corresponde a la voz de un hombre zuliano, sería una decisión errada, aun cuando el 85.714% de las palabras indican que se trata de un hombre y el 57.14% indican que es del Zulia.

La situación anterior se puede presentar cuando no se obtiene al menos el 50% de reconocimiento de las palabras presentes en la pronunciación de entrada (en la salida HTK).

8.6.2. Resultados de las pruebas de los cinco dialectos

Con respecto a las pruebas donde se le pedía al reconocedor que discriminara entre los cien modelos de los dígitos, más el silencio, los resultados obtenidos cuando se trabajaba con archivos de entrenamiento fueron cercanos al 100% en dialecto y género, pero cuando se trabajaba con archivos de test no se lograba prácticamente ningún tipo de reconocimiento a nivel de dialectos, mientras que a nivel de género el porcentaje era alto.

A continuación se muestra una salida típica del último caso.

```
---------------------- Overall Results -------------------------
SENT: %Correct=2.90 [H=4, S=134, N=138]
WORD: %Corr=23.26, Acc=7.33 [H=181, D=12 S=585, I=124, N=778]
```

En este caso se obtuvieron salidas de la forma siguiente:

Salida esperada: acerof adosf acuatrof acincof anuevef aochof asietef anuevef acerof
Salida obtenida: zcerof zdosf ccuatrof llcincoh onuevef aochof csieteh llnuevef llcerof

cuando se trabajaba con una gramática que admitía, que cualquier palabra podía ser seguida por cualquier otra.

Se puede observar que el reconocedor falla en el reconocimiento de la secuencia, pues sólo identificó como correcta una palabra (aochof) de las 9 que contiene la secuencia. Sin embargo, todas las palabras que presenta como salidas se refieren al dígito correcto, y la mayoría de esas palabras indican que se trata de la pronunciación de una mujer.

El problema que se presenta en estas pruebas es que el reconocedor no puede determinar la zona dialectal a la que pertenece el locutor (esto se refleja en la salida HTK que se muestra arriba), pero si lo hace en cuanto al género y en cuanto al dígito.

Haciendo la revisión salida por salida, de los resultados de esta prueba se encontró que el porcentaje de reconocimiento a nivel de género fue de 97.1%, y si sólo nos fijamos en el dígito y no en el dialecto, el nivel de reconocimiento a nivel de los dígitos es aproximadamente 90%, mientras que a nivel de dialectos, obviamente es bajo, aproximadamente 18.31%.

A continuación se presenta otro resultado, donde se trabajó con los archivos de test, pero con la gramática que aparece en el anexo D como **gramáticatotal1.**

```
----------------------- Overall Results -------------------------
SENT: %Correct=0.72 [H=1, S=137, N=138]
WORD: %Corr=34.89, Acc=18.92 [H=389, D=54, S=672, I=178, N=1115]
=============================================================
```

En el anexo D se muestra parte de las salidas de esta prueba en **salidadialectostest4**, donde se puede observar que es imposible distinguir los dialectos, pero si es posible determinar el género del hablante en un alto porcentaje.

8.6.3. Resultados de las pruebas de los tres dialectos

Después de revisar los resultados obtenidos en las pruebas donde se trabajó con cinco zonas dialectales, donde se observó que todos los dialectos se confundían, y que era marcado el porcentaje de confusión con el centro y con los llanos, se procedió a construir un reconocedor que manejara sólo los dialectos de los Andes, el Zulia y Oriente.

A continuación se presenta una salida de una de las pruebas de este tipo, donde se re-estimaron los modelos 36 veces y se utilizaron archivos de test.

```
---------------------------------------- Overall Results ----------------------------------
SENT: %Correct=7.69 [H=8, S=96, N=104]
WORD: %Corr=42.57, Acc=32.09 [H=252, D=15, S=325, I=62, N=592]
=============================================================
```

Luego de hacer la revisión salida por salida de los resultados de esta prueba, se determinó que el porcentaje de reconocimiento a nivel de género era de 99.04%, mientras que a nivel de dialectos era de 35.57%.

8.6.4. Otras pruebas

También se realizaron pruebas donde intervenían cuatro dialectos, como fue el caso donde se entrenó un reconocedor que tratara con el dialecto zuliano, andino, llanero y oriental. Se dejó por fuera el dialecto central, porque por algún momento se llegó a considerar que parecía el dialecto que generaba mayor confusión al reconocedor de cinco dialectos.

Como en todas las pruebas, para este reconocedor se re-estimaron los modelos diferentes números de veces, para detectar posible sobre-entrenamiento.

Los mejores resultados encontrados fueron un porcentaje de reconocimiento del 97.46% para el género y un 25.42% para los dialectos. Se trabajó con un total de 118 archivos de test.

De la misma manera se entrenaron reconocedores para distinguir el género y los dialectos andino y central. Los mejores resultados fueron 100% género y 40.98% dialecto de un total de 61 archivos de test.

En cuanto al mejor reconocedor de dialectos zuliano y llanero, los mejores resultados fueron 100% género y 46.2% dialectos de un total de 52 archivos de test.

8.7. ANÁLISIS DE RESULTADOS Y CONCLUSIONES

Los resultados obtenidos en cada una de las pruebas descritas indican que bajo estas condiciones, donde se modelan palabras, y donde el tipo de palabras son los dígitos, es imposible crear reconocedores que puedan distinguir los dialectos venezolanos que se trataron en este experimento. Sin embargo, lo rescatable de estas pruebas es que se observó que puede ser completamente factible crear reconocedores con modelos de palabras que puedan distinguir, cuándo habla un hombre de cuándo habla una mujer.

Los resultados obtenidos en estas pruebas, nos llevaron a realizar otro tipo de experimentos, donde se elaboraron otros tipos de modelos que pueden ser la base para un sistema de reconocimiento de dialectos de mejor rendimiento. Los resultados de esos otros experimentos se explican en el próximo capítulo.

CAPITULO IX

RECONOCIMIENTO AUTOMÁTICO DE DIALECTOS VENEZOLANOS POR MEDIO DE MODELOS DE FONOS

9.1. INTRODUCCIÓN

En este capítulo se muestran los resultados obtenidos de un conjunto de pruebas en las que se crearon una serie de modelos de fonos, a través de los cuales se pretendía analizar la capacidad que podría tener un sistema de reconocimiento, para determinar a qué zona dialectal venezolana pertenecía el locutor de cada pronunciación que se le daba como entrada. Para este tipo de pruebas, se trabajó con las cinco zonas dialectales conocidas: zuliana, andina, central, llanera y sud-oriental.

Con el fin de averiguar si sería posible dotar a una máquina con la capacidad de distinguir esos cinco dialectos venezolanos, se crearon modelos de fonos que incluyeran las características de la voz de hombres y de mujeres de cada zona dialectal. En este caso, sólo se le pedía al reconocedor que distinguiera los dialectos y no el género del locutor como adicionalmente, se requería en el trabajo descrito en el capítulo precedente; ésta es la razón por la cual se trabajó con modelos globales por región dialectal.

9.2. JUSTIFICACIÓN DE ESTAS PRUEBAS

Cuando se realizó el reconocimiento simultáneo de los dialectos y del género del locutor utilizando modelos de palabras (dígitos), los resultados de las pruebas indicaron que esa forma de reconocimiento funciona perfectamente para distinguir sí el locutor es un hombre o una mujer; mientras que no permite distinguir en forma eficiente el dialecto del locutor. Esto llevó a considerar el modelado de otras unidades (fonos) del habla venezolana, donde el único objetivo era buscar una forma adecuada, para lograr un nivel aceptable de reconocimiento de los dialectos.

9.3. LOS MODELOS UTILIZADOS

Los modelos de los fonos que se utilizaron fueron los siguientes: s, E, r, o, u, n, d, s2, t, e, k, w, a, i, j, ts, b, sil y sp. Se trata de los distintos sonidos (fonos) que se encontraron en un conjunto de secuencias de pronunciaciones de dígitos pertenecientes a la SPEECHDAT Venezolana.

Sil y sp son los símbolos que se utilizaron para representar el silencio existente al principio y al final de cada grabación y para los espacios entre palabras.

La representación mediante símbolos de los sonidos que se encuentran en las pronunciaciones de los dígitos venezolanos, es una versión del alfabeto fonético SAMPA [52], como se puede apreciar en la tabla 9.1.

Para esta forma de reconocimiento de los dialectos venezolanos, cuando se hizo el etiquetado, además de realizar la transcripción fonética a cada sonido, se agregó al símbolo fonético, otro símbolo constituido por una letra que indicaba la región a la que pertenecía el locutor. La nomenclatura utilizada fue la siguiente: andinos (a), zulianos (z), centrales (c), llaneros (l) y orientales (o).

Así por ejemplo, el dígito "cuatro", cuya transcripción SAMPA sería "k w a t r o"; si era pronunciado por un andino, se transcribiría como "ak aw aa at ar ao".

Ese tipo de nomenclatura no se aplica al sonido "sp" ni al "sil", ya que estos dos tipos de sonidos son independientes de la zona dialectal.

Bajo la forma descrita, para representar los sonidos presentes en los dígitos pronunciados por personas de las cinco zonas dialectales, se tenían entonces: 17 modelos diferentes de los sonidos de cada región, más los dos modelos para representar el silencio y las pausas entre las palabras. Esto daba lugar a un conjunto de 87 modelos (17*5+2).

Tabla 9.1. Transcripción fonética venezolana para los dígitos en una versión SAMPA.

Dígito	Transcripción fonética (SAMPA)
CERO	[sEro]
UNO	[uno]
DOS	[dos2]
TRES	[tres2]
CUATRO	[kwatro]
CINCO	[sinko]
SEIS	[sejs2]
SIETE	[sjEte]
OCHO	[otso]
NUEVE	[nwEbe]

El procedimiento que se siguió para etiquetar cada fono, fue el mismo explicado en el capítulo anterior para etiquetar los modelos de los dígitos por regiones, sólo que en ese caso, cada dígito generaba un modelo, mientras que ahora teníamos varios modelos por dígito.

Los modelos que se crearon para estas pruebas fueron MOM tipo izquierda-derecha, de 5 estados, se varió el número de mezclas gausianas por estado, desde 1 a 12, y se hicieron entre 5 y 36 re-estimaciones. Los mejores resultados se obtuvieron con 7 mezclas gausianas y con 11 re-estimaciones por modelo.

9.4. EL DICCIONARIO UTILIZADO

El diccionario utilizado para el reconocimiento de los dialectos por medio de modelos de fonos, fue el siguiente:

acero as aE ar ao

auno	au an ao
ados	ad ao as2
atres	at ar ae as2
acuatro	ak aw at ar ao
.	.
.	.
.	.
zseis	zs ze zj zs2
zsiete	zs zj zE zt ze
zocho	zo zts zo
znueve	zn zw zE zb ze
sil	sil
sp	sp

Los tres puntos intercalados en el diccionario indican que ahí están incluidos también, los modelos de las zonas llanera, sudoriental y central.

9.5. CONSTRUCCIÓN DE LOS MODELOS

Para la construcción de los modelos se utilizó un conjunto de pronunciaciones de secuencias de dígitos de hombres y de mujeres de cada zona dialectal, básicamente los mismos archivos que se emplearon para el tipo de reconocimiento que se describió en el capítulo precedente.

La parametrización de los archivos y la estimación de los modelos, siguieron exactamente el mismo procedimiento que el efectuado para hacer reconocimiento de dialectos con modelos de palabras.

9.6. PRUEBAS REALIZADAS

Las pruebas consistieron en presentarle a los reconocedores señales que contenían pronunciaciones de dígitos en secuencia, y se esperaba que de la respuesta, se pudiera obtener una medida de su capacidad para indicar a qué zona dialectal pertenecía ese tipo de voz.

Las pruebas de reconocimiento del dialecto con modelos de fonos, se realizaron de la siguiente manera:

9.6.1. Reconocimiento de dos dialectos: se crearon los modelos de la región andina y los modelos de la región zuliana.

Se trabajó con 60 archivos de entrenamiento, 30 del Zulia y 30 de la región andina.

A continuación se presenta la salida HTK para los archivos de entrenamiento.

```
------------------------ Overall Results --------------------------
SENT: %Correct=70.00 [H=42, S=18, N=60]
WORD: %Corr=97.75, Acc=97.30 [H=651, D=7, S=8, I=3, N=666]
================================================================
```

La salida HTK para una prueba con archivos de test es la siguiente:


```
------------------------ Overall Results --------------------------
SENT: %Correct=22.22 [H=2, S=7, N=9]
WORD: %Corr=83.33, Acc=78.43 [H=85, D=9, S=8, I=5, N=102]
================================================================
```

9.6.2. Reconocimiento de tres dialectos: se crearon los modelos de las regiones andina, zuliana y central.

Se trabajó con 96 archivos de entrenamiento, 30 del Zulia, 30 de la región andina y 36 de la región central.

A continuación se presenta la salida HTK para los archivos de entrenamiento.

```
------------------------ Overall Results --------------------------
SENT: %Correct=75 [H=72, S=24, N=96]
WORD: %Corr=98.04, Acc=97.35 [H=998, D=8, S=12, I=7, N=1018]
================================================================
```

La salida HTK para una prueba con archivos de test es la siguiente:

```
---------------------- Overall Results ----------------------
SENT: %Correct=6.25 [H=1, S=15, N=16]
WORD: %Corr=77.51, Acc=68.05 [H=131, D=0, S=38, I=16, N=169]
================================================================
```

9.6.3. Reconocimiento de cuatro dialectos: se crearon los modelos de las regiones andina, zuliana, central y llanera.

Se trabajó con 122 archivos de entrenamiento, 30 del Zulia, 30 de la región andina, 36 de la región central y 26 de la región llanera.

A continuación se presenta la salida HTK para los archivos de entrenamiento.

```
---------------------- Overall Results ----------------------
SENT: %Correct=72.95 [H=89, S=33, N=122]
WORD: %Corr=98.14, Acc=96.74 [H=1323, D=10, S=15, I=19, N=1348]
================================================================
```


La salida HTK para una prueba con archivos de test es la siguiente:

```
---------------------- Overall Results ----------------------
SENT: %Correct=9.09 [H=2, S=20, N=22]
WORD: %Corr=70.72, Acc=60.36 [H=157, D=3, S=62, I=23, N=222]
================================================================
```

9.6.4. Reconocimiento de cinco dialectos: se crearon modelos de las cinco zonas dialectales. En este caso, el número de archivos de entrenamiento fue de 153.

A continuación se presenta la salida HTK para los archivos de entrenamiento:

```
---------------------- Overall Results -------------------------
SENT: %Correct=69,28 [H=106, S=47, N=153]
WORD: %Corr=98.04, Acc=96.39 [H=1656, D=14, S=19, I=28, N=1689]
============================================================
```

La salida HTK para una prueba con archivos de test es la siguiente:

```
---------------------- Overall Results -------------------------
SENT: %Correct=3.45 [H=1, S=28, N=29]
WORD: %Corr=62.63, Acc=45.89 [H=184, D=7, S=101, I=50, N=292]
============================================================
```

9.7. RESUMEN DE LOS RESULTADOS

El criterio que se siguió para medir la capacidad del reconocedor de dialectos, fue parecido al que se explicó, cuando se trabajó con modelos de los dígitos, sólo que en este caso, se calculaba un solo porcentaje (el porcentaje de dialectos), sobre la secuencia de palabras que presentaba el reconocedor como respuesta.

Ese porcentaje se calculaba de la siguiente manera: supóngase que se tiene la respuesta que se desearía que diera el reconocedor y la respuesta que éste daba, en una prueba cualquiera.

Salida esperada: sil cdos ctres ccero cocho csiete csiete ctres sp ccinco sil
Salida obtenida: sil cdos atres ccero zocho sp csiete ctres sp ccinco sil

El porcentaje de reconocimiento se calculaba contando sólo las palabras de los dígitos (no los silencios ni las pausas). En este ejemplo, el reconocedor presenta siete palabras de dígitos (en la salida obtenida), de las cuales 5 palabras (71.43%) indican que el dialecto es de una persona de la zona central, 1 palabra (14.28%) indica que se trata de una persona andina y 1 palabra (14.28%) indica que se trata de una persona zuliana.

Se consideraba que el reconocedor identificaba correctamente el dialecto al que correspondía la secuencia pronunciada, si el porcentaje calculado de la manera señalada, superaba el 60 por

ciento de las palabras de las secuencias (este porcentaje umbral es perfectamente válido cuando el nivel de reconocimiento a nivel de palabras que presenta HTK, es superior a 50%).

Por lo tanto, para el ejemplo que se acaba de presentar, el 71.43% de las palabras indican que la voz pertenece a una persona central, como lo es efectivamente, tal como se observa en la salida esperada.

Es importante tomar en cuenta que la forma descrita para obtener los porcentajes de acierto, arrojan valores distintos a los que produce HTK en las salidas mostradas de cada prueba. Sin embargo, el nivel de reconocimiento por palabras que da HTK es en este caso, una buena aproximación del nivel de reconocimiento de los dialectos.

Después de realizar los cálculos de los porcentajes (en las pruebas, donde las salidas HTK presentaban los porcentajes más altos), por cada secuencia de salida, en cada una de las pruebas señaladas, se encontró que el mejor porcentaje de reconocimiento para 5 dialectos fue de 52%, para la prueba de 4 dialectos un 64%, para la prueba de 3 dialectos un 83.4% y para la prueba de 2 dialectos un 84.21%.

Todos esos porcentajes, se refieren al reconocimiento con señales de voz de test, mientras que con señales de entrenamiento, el reconocimiento en la mayoría de los casos resultó de 100%.

9.8. ANÁLISIS DE LOS RESULTADOS DE LAS PRUEBAS

Como se puede observar, bajo esta forma de trabajo, los resultados de la identificación de los dialectos mejoran en cantidades considerables, respecto a las pruebas que se describieron en el capítulo anterior.

9.9. CONCLUSIONES DE LAS PRUEBAS

Es muy difícil dotar máquinas con una capacidad, suficientemente buena, de distinguir los dialectos venezolanos descritos, cuando se parte de modelos de pronunciaciones de dígitos leídos.

Los modelos de los fonos venezolanos producen mejores resultados que los modelos de las palabras para el reconocimiento de los dialectos.

Las pronunciaciones de secuencias de los dígitos de los venezolanos capturan, cuando son leídos, muy poco de las características regionales de la voz, por lo que en estudios posteriores se debe trabajar con modelos de fonos que se construyan en base a pronunciaciones de habla espontánea o en base a palabras internas de frases leídas.

Entre menor sea el número de propiedades que se le exija a un reconocedor, mayor será su capacidad de reconocimiento. Por ejemplo, en el capítulo anterior se tenían modelos a través de los cuales se reconocían palabras (los dígitos), el género del locutor y la zona dialectal a la que pertenecía dicho locutor; allí los resultados no fueron ni siquiera aceptables en el caso del dialecto, mientras que en las pruebas que se describieron en este capítulo, se utilizaron modelos a través de los cuales se reconocían sólo las palabras y el dialecto, con lo cuales los resultados mejoraron considerablemente.

CAPITULO X

RECONOCIMIENTO AUTOMÁTICO DE SECUENCIAS DE DIGITOS POR MEDIO DE MODELOS DE FONOS

10.1. INTRODUCCIÓN

En este capítulo se muestran los resultados obtenidos de dos tipos de pruebas en las que se realizó reconocimiento automático de secuencias de dígitos pronunciados por múltiples hablantes del territorio nacional. En estas pruebas, se trabajó con modelos de fonos, a diferencia del caso mostrado en el capítulo 5, donde se presenta igual que aquí, el reconocimiento de cadenas de dígitos pero trabajando con un modelo por cada dígito.

El objetivo de las pruebas que se describen en este capítulo era averiguar, por un lado, qué modelos (fonos o palabras) resultan más robustos para hacer reconocimiento de secuencias de dígitos venezolanos y, por otro lado, averiguar cómo era el nivel de reconocimiento cuando se obtenían los modelos, partiendo de realizaciones de fonos obtenidos de pronunciaciones patrones de otras palabras que no fueran precisamente dígitos.

10.2. JUSTIFICACIÓN DE ESTAS PRUEBAS

Aun cuando los resultados obtenidos en el reconocimiento de pronunciaciones de secuencias de dígitos, explicados en el capítulo 5, se pueden considerar como satisfactorios, es claro que en aplicaciones reales se requiere un porcentaje de reconocimiento muy cercano al cien por ciento, para que este tipo de tecnología pueda ser aceptada. Es decir, el grado de confiabilidad que se desea en las aplicaciones reales es tal, que no hay margen para el error, pues sólo pensar que la máquina no reconozca completamente, por ejemplo, la secuencia de dígitos de nuestra cédula, lleva a dudar de la verdadera capacidad de esos sistemas.

Es por eso que se planteó utilizar otro tipo de modelos, en este caso modelos de fonos para tratar de mejorar los resultados obtenidos en aquellos experimentos.

Para este conjunto de pruebas no fue necesario construir más modelos de los fonos, debido a que se utilizaron los mismos modelos que fueron creados para los tipos de pruebas que se describieron en los capítulos 7 y 9.

10.3. LOS MODELOS UTILIZADOS

Se trabajó con dos grupos de símbolos para representar los fonos: un conjunto consistió en los símbolos que se utilizaron en las pruebas descritas en el capítulo 7 y el otro conjunto, fue el que se utilizó en las pruebas descritas en el capítulo 9.

La razón por la cual aparecen estos dos conjuntos de símbolos, es que no se cuenta todavía con una forma específica o uniforme para representar los sonidos del habla venezolana para hacer este tipo de pruebas, por lo que se recurre al uso de variantes del alfabeto SAMPA [52].

A continuación se muestran los dos conjuntos de símbolos:

Conjunto 1: a, b, B, c, d, D, e, f, g, G, h, i, j, k, l, m, n, N, M, o, p, r, R, s, t, u, w, y y sil.

Este conjunto 1, representa el conjunto de todos los sonidos encontrados en los archivos de voz que se utilizaron en el reconocimiento de las fechas venezolanas. Se podría indicar que este conjunto de fonos es general y bastante cercano al conjunto de los fonos distintos del habla venezolana, debido a que en las fechas intervienen una gran cantidad de palabras de nuestro lenguaje y por lo tanto, una buena cantidad de realizaciones de los sonidos del habla de Venezuela.

Conjunto 2: s, E, r, o, u, n, d, s2, t, e, k, w, a, i, j, ts, b, sil y sp.

Este conjunto 2, representa sólo los fonos presentes en pronunciaciones venezolanas de los dígitos.

10.4. EL DICCIONARIO UTILIZADO

El diccionario utilizado para este reconocimiento, por medio de los modelos obtenidos a partir de las pronunciaciones de dígitos correspondientes a las zonas dialectales (capítulo 9), tiene la siguiente forma por cada dígito:

cero	as aE ar ao
uno	au an ao
dos	ad ao as2
tres	at ar ae as2
cuatro	ak aw at ar ao
cinco	as ai an ak ao
seis	as ae aj as2
siete	as aj aE at ae
ocho	ao ats ao
nueve	an aw aE ab ae
.	.
.	.
.	.
seis	zs ze zj zs2
siete	zs zj zE zt ze
ocho	zo zts zo
nueve	zn zw zE zb ze
sil	sil
sp	sp

Los puntos que aparecen en el diccionario, indican que están incluidos los modelos de las cinco zonas dialectales.

La diferencia de este diccionario, respecto al diccionario empleado en el reconocimiento de los dialectos, es que aquí cada salida del diccionario tiene cinco secuencias alternativas de fonos, mientras que en aquél caso, cada salida tenía asociada una secuencia.

Por ejemplo, en este diccionario la palabra cuatro como salida, se obtiene de las cinco formas siguientes:

cuatro ak aw aa at ar ao
cuatro zk zw za zt zr zo
cuatro ck cw ca ct cr co
cuatro lk lw la lt lr lo
cuatro ok ow oa ot or oo

Así como se puede obtener la palabra cuatro, cada dígito puede ser obtenido de diversas formas: un dígito puede ser obtenido con modelos de fonos creados a partir de realizaciones de dígitos del habla de personas del Zulia, o de personas de los Andes, o de personas de la región llanera o de la región central o de la región sudoriental. Por esa razón, es que en el diccionario aparecen cinco formas posibles de producir cada dígito.

El diccionario utilizado para el reconocimiento de los dígitos por medio de los modelos obtenidos a partir de las fechas, tiene la siguiente forma por cada dígito:

Cero s e r o
Uno u n o
Dos d o s
Dos d o h
Dos D o s
Dos D o h
Dos do
Dos Do
Tres t r e s
Tres t r e
Tres t r e h
Cuatro k w a t r o
Cinco s i N k o
Seis s e j s

Seis s e j

Siete s j e t e

Ocho o c o

Nueve n w e b e

Se puede observar que se utilizaron para algunos dígitos transcripciones alternativas, que se emplean debido al conocimiento que se tiene de cómo los venezolanos pronuncian los dígitos.

10.5. PRUEBAS REALIZADAS

Las pruebas de reconocimiento de pronunciaciones de secuencias de dígitos con los tipos de modelos descritos, fueron las siguientes:

10.5.1. Reconocimiento de secuencias de dígitos, utilizando los modelos de fonos creados a partir de pronunciaciones de fechas

Estas pruebas consistieron en presentarle como entrada al reconocedor, el mismo conjunto de secuencias de pronunciaciones de dígitos pertenecientes al corpus de test que se utilizó en las pruebas descritas en el capítulo 5, recordemos que en esa oportunidad se estaba haciendo reconocimiento de secuencias de dígitos pero trabajando con modelos de palabras (un modelo por dígito).

Los resultados se muestran a continuación.

```
----------------------- Overall Results --------------------------
SENT: %Correct=10.32 [H=13, S=113, N=126]
WORD: %Corr=83.08, Acc=45.35 [H=599, D=2, S=120, I=272, N=721]
=================================================================
```

10.5.2. Reconocimiento de secuencias de dígitos, utilizando los modelos de fonos creados a partir de pronunciaciones de dígitos de las cinco zonas dialectales

Estas pruebas consistieron en presentarle como entrada al reconocedor, un conjunto de secuencias de pronunciaciones de dígitos pertenecientes al corpus de test que se utilizó para reconocer dialectos a través de fonos.

A continuación se presenta la salida HTK de una de estas pruebas:

```
---------------------- Overall Results -------------------------
SENT: %Correct=82.75 [H=24, S=5, N=29]
WORD: %Corr=99.31, Acc=96.23 [H=290, D=0, S=2, I=9, N=292]
==========================================================
```

10.6. ANÁLISIS DE LOS RESULTADOS DE LAS PRUEBAS

Los resultados obtenidos en el primer tipo de pruebas muestran claramente una disminución en el nivel de reconocimiento, con respecto al obtenido cuando se realizó este tipo de prueba donde se trabajó con modelos de los dígitos. En esa oportunidad se obtuvo un reconocimiento a nivel de dígitos del 98.47% y a nivel de secuencias completas un 64.29% para el corpus de test.

En este caso, en el que se trabajó con modelos de fonos que eran independientes del contexto (esos modelos no fueron construidos de pronunciaciones tomadas de dígitos únicamente, sino que fueron obtenidos de pronunciaciones de otras palabras que intervienen en la pronunciación de fechas), los resultados fueron decepcionantes, en el sentido de que con estas pruebas se deseaba averiguar sí sería posible construir modelos generales de fonos, que se pudieran utilizar para reconocer palabras y oraciones distintas a fechas, y no se logró ese objetivo, aun cuando las secuencias a reconocer contenían sólo dígitos.

Por otro lado, los resultados obtenidos partiendo de modelos de fonos obtenidos a partir de pronunciaciones de dígitos, son excelentes.

10.7. CONCLUSIONES DE LAS PRUEBAS

Es muy difícil crear modelos de voz generales, que sirvan para reconocer cualquier palabra. Por lo tanto, para cada aplicación para la cual se construya un reconocedor, sus modelos de voz deben obtenerse a partir de realizaciones de palabras y oraciones propias de la aplicación, con el fin de obtener buenos niveles de reconocimiento.

Con modelos de fonos se logran resultados superiores, a cuando se trabaja con modelos de palabras, en casos donde se hace reconocimiento de secuencias de dígitos.

En aplicaciones donde se trabaje con grandes vocabularios resultará mas beneficioso desde el punto de vista del espacio de memoria requerido, así como del tiempo y los datos invertidos en el entrenamiento, usar modelos de fonos que podrían llegar a ser de menos de cien modelos, en muchos casos, y no de tantos como palabras tenga el vocabulario del reconocedor.

CAPITULO XI

SISTEMA INCREMENTAL GENERADOR DE ORACIONES Y DE DESCODIFICACIÓN LINGÜÍSTICA

11.1. INTRODUCCIÓN

En este capítulo se describe la implementación experimental de un sistema que tiene las siguientes propiedades: permite crear modelos del lenguaje para reconocedores automáticos del habla, genera oraciones a partir de esos modelos y hace descodificación lingüística de secuencias de palabras. Se supone que las secuencias de palabras son suministradas por un módulo de descodificación acústica.

11.2. JUSTIFICACIÓN DEL DESARROLLO DE ESTE TIPO DE SISTEMAS

El creciente e indetenible desarrollo en el campo de la Tecnología del Habla, ha llevado a pensar que puede ser posible el diseño y construcción de sistemas automáticos de reconocimiento del habla, cuyo módulo descodificador linguístico comprenda un modelado de lenguaje que copie y codifique todos los componentes gramaticales desde un corpus de entrenamiento [31][33][39][56].

Por nuestro lado, como parte de un intento por incursionar en este apasionante mundo del desarrollo de ese componente tan importante de los sistemas de reconocimiento, como es, el subsistema de descodificación linguística, hemos desarrollado un sistema (un conjunto de programas de computación), a través del cual se ha observado que partiendo del modelo (de lenguaje) obtenido a partir de un pequeño corpus compuesto de un conjunto de oraciones y párrafos de un contexto particular, se puede generar automáticamente una gran cantidad de oraciones gramaticalmente válidas y de ese contexto. También, el sistema puede realizar pruebas de descodificación lingüística donde rechaza como fuera de contexto o incorrectas

gramaticalmente, aquellas secuencias de entrada que contiene palabras que no están presentes en su vocabulario o que contienen historias que no tiene codificadas en su memoria.

Pensamos, que un tipo de sistema de descodificación y tratamiento de la información como el que se presenta en este capítulo, puede cubrir de manera exitosa el reconocimiento en diversos contextos, donde el tamaño del vocabulario sea de varios miles de palabras.

La construcción de este sistema tiene la finalidad de contribuir en el futuro, al desarrollo de un sistema de reconocimiento producto de la Universidad de Los Andes, por esa razón se han programado diversos módulos, aparte del que se está describiendo, como son: módulos de parametrización por análisis de predicción lineal y cepstral, dos módulos para hacer cuantificación vectorial a través del algoritmo LBG y LVQ [1][62], un módulo para crear y probar MOM de observaciones discretas a través de los algoritmos de Viterbi y Baum-Welch (para estimar los parámetros de los MOM y obtener la capacidad de producción de observaciones, en los dos casos) y un módulo para hacer adquisición de la voz a través de micrófonos. Recordemos que la mayoría de las pruebas que se presentan en está tesis, a excepción de la que se está describiendo, y la explicada en el capítulo 3, se realizaron con HTK, lo que significa que de esos estudios sólo se pueden preservar los modelos creados, y por su puesto la experiencia adquirida en cuanto al diseño de reconocedores de aplicaciones específicas, puesto que si se desea construir un reconocedor autónomo de desarrollo propio, se tienen que implementar la mayoría de los algoritmos necesarios para manipular los modelos que fueron creados y probados con HTK.

11.3. TERMINOLOGÍA UTILIZADA

Corpus de entrenamiento: el conjunto formado por las oraciones y párrafos a partir del cual se construye el modelo de un contexto o lenguaje de una aplicación, tanto para el reconocimiento como para la generación de las oraciones.

Vocabulario: el vocabulario comprende el conjunto de palabras distintas que se encuentran en el corpus de entrenamiento.

Entrenamiento: proceso mediante el cual se crea el modelo del contexto o de lenguaje.

Historias: conjuntos de palabras que aparecen en forma contigua en el corpus de entrenamiento [22][56].

Ejemplo de historias: sea la siguiente, una oración presente en el corpus de entrenamiento: "tres razones parecen ser el origen de este hecho".

Una historia de dos palabras sería: "el origen".

Una historia de tres palabras sería: "el origen de".

Contexto: área de aplicación a la cual pertenecen las oraciones y párrafos que componen el corpus de entrenamiento.

Reconocimiento: el proceso de descodificación de secuencias de palabras que se le presentan al sistema, a través del cual se determina si dicha secuencia es una oración válida respecto a las reglas gramaticales presentes en el contexto que se modela.

Generación de oraciones: proceso mediante el cual se crea una oración a partir del modelo del contexto.

Oración gramáticalmente válida: oración que tiene una estructura que sigue las reglas gramaticales encontradas en el corpus de entrenamiento.

Hipótesis de oraciones: conjunto de posibles oraciones a las que podría corresponder una secuencia de palabras a reconocer o a generar.

Reglas gramaticales: la combinación de las palabras que se encuentran en el corpus de entrenamiento.

11.4. ARQUITECTURA DEL SISTEMA

En la figura 11.1 se muestran los elementos principales del sistema, sus entradas y salidas, y se da una idea gráfica de cómo interactúan dichos elementos.

Se puede apreciar que el sistema desarrollado comprende un módulo que se encarga de generar el modelo del contexto (generador de modelo), a partir de un corpus de entrenamiento, un módulo que se encarga de producir oraciones a partir del modelo del contexto y del vocabulario del sistema (el generador de oraciones), y un módulo que se encarga de determinar si una secuencia de palabras es válida para el lenguaje de la aplicación (el descodificador lingüístico), para lo cual se utiliza también el modelo del contexto y el vocabulario.

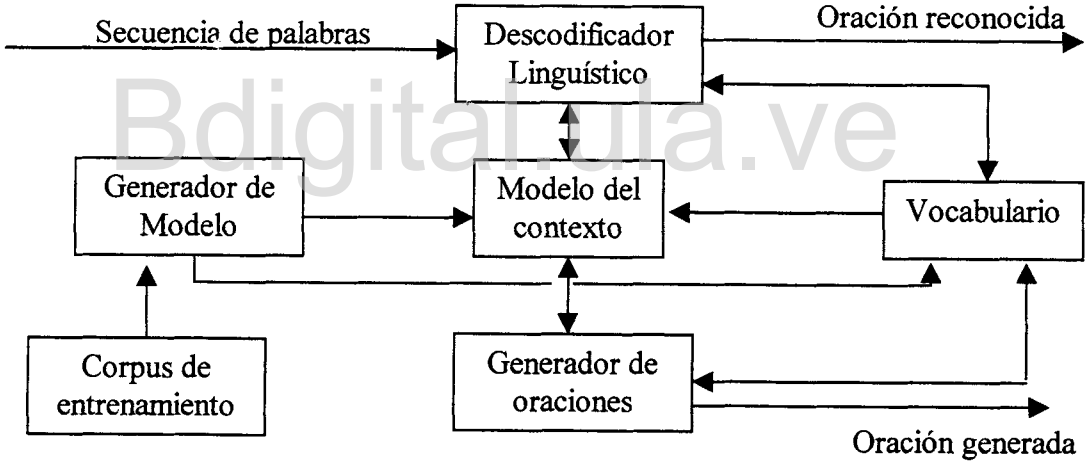


Figura 11.1. Estructura del sistema

Este sistema tiene la particularidad, de que permite crear el descodificador lingüístico en una forma completamente independiente de la construcción del descodificador acústico, a diferencia de muchos trabajos descritos en [56]. Su construcción supone que el descodificador acústico se encarga de detectar los sonidos básicos del habla y que con estos construye las palabras, es decir, el tipo de descodificador lingüístico desarrollado aquí, recibiría secuencias de palabras por parte del descodificador acústico. El hecho de que reciba secuencias de palabras, es otra diferencia con otros sistemas existentes [56], debido a que esta forma de trabajo requiere que el descodificador acústico necesariamente esté dotado de un módulo de

análisis de léxico. Otra propiedad que se le puede atribuir al sistema es que permite averiguar fuera de línea la cantidad aproximada de oraciones que puede manejar para una aplicación dada, lo cual se logra a través del módulo Generador de oraciones, esto último también, en forma independiente del descodificador acústico.

11.5. GENERADOR DE MODELOS DE CONTEXTOS. ALGORITMO PROPUESTO.

Para crear el modelo se parte de un conjunto gramaticalmente correcto de oraciones y párrafos propio del contexto que se quiere modelar. El entrenamiento consiste básicamente en la búsqueda, codificación y almacenamiento de la ocurrencia de historias de palabras contiguas dentro de las oraciones y párrafos del corpus de entrenamiento.

Se crean bloques codificados de historias de palabras. Las palabras presentes en las historias se codifican a través de números enteros.

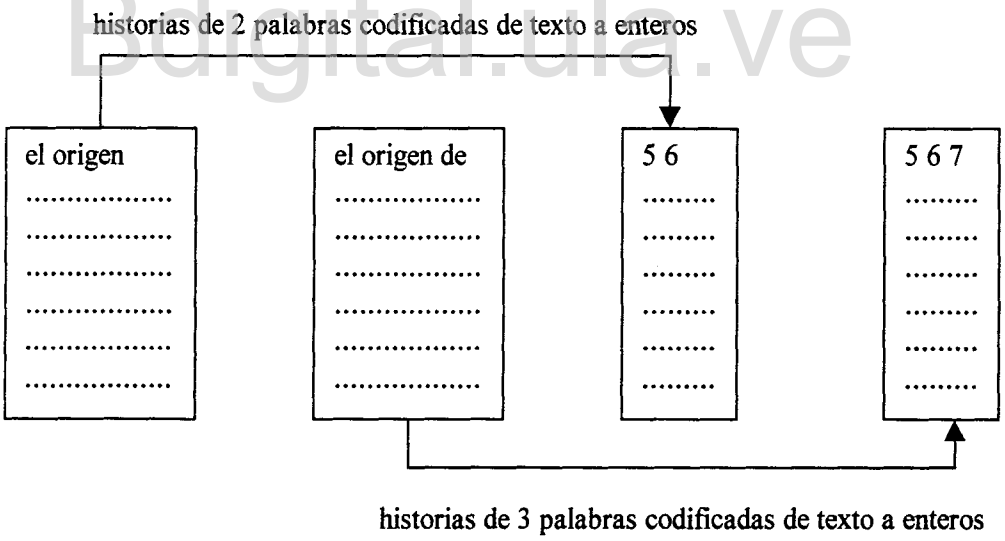


Figura 11.2: Ejemplos de bloques de historias tomadas del corpus de entrenamiento

En la figura 11.2, se da una idea de cómo se ubican las historias presentes en el corpus de entrenamiento. Allí, se puede observar que la historia "el origen" se codifica como la secuencia de enteros " 5 6" y la historia "el origen de" se codifica como la secuencia de enteros "5 6 7" (esos números corresponden a la posición que ocupan las palabras en la

oración dada en la sección 11.3, donde se definen las historias). El número entero se asigna de acuerdo a la posición que tiene cada una de las palabras en el vocabulario obtenido del corpus.

A la primera palabra que aparece en el primer corpus de entrenamiento se le asigna el número 1 (se puede trabajar de manera incremental con varios corpus), luego, a la siguiente palabra distinta a la primera, se le asigna el número 2 y así sucesivamente. Entonces, a la n-ésima palabra distinta del corpus, se le asigna el número n.

Se forman los bloques siguientes:

Un bloque que contiene una lista codificada de las palabras que en el corpus inician las oraciones; éste bloque es usado por el generador de oraciones y se identifica como bloque1.

Un bloque que contiene una lista de historias de dos palabras. Este bloque puede ser usado tanto por el generador de oraciones como por el descodificador linguístico. A éste se le identifica como bloque2.

Un bloque que contiene una lista de historias de tres palabras, igual que el bloque2, puede ser usado tanto por el generador como por el descodificador, al que se le identifica como bloque3.

Un bloque que contiene una lista de historias de tres palabras como el bloque3, pero a diferencia de aquel, cada historia contiene las tres últimas palabras de cada oración y párrafo del corpus de entrenamiento. Este bloque le permite al generador de oraciones darse cuenta de cuándo ha de finalizar la producción de una oración. Se trata del bloque4.

La codificación del corpus de entrenamiento en los bloques 1, 2, 3 y 4, se hace para facilitar la creación de las hipótesis de las oraciones para el reconocimiento linguístico, para la generación de oraciones y para agilizar el manejo de los datos.

Una vez que se crean y se codifican los bloques antes mencionados, se descartan ciertas historias para bajar un poco la redundancia en la codificación, y para agilizar el trabajo de los programas que reconocen y generan oraciones. La eliminación de historias es un tanto

arbitraria; por ejemplo, se hicieron pruebas donde se eliminan por un lado, las historias de dos palabras cuya frecuencia de aparición en el corpus es baja, y por otro lado, se almacenan las historias de tres palabras que se inician con aquellas historias eliminadas, de dos palabras.

Por ejemplo, supóngase que en el corpus, el par de palabras "la linguística" aparece 7 veces y el par "sujeto a" aparece 1 vez. Entonces, se almacena el primer par, mientras que el segundo par no, pero en su lugar, se almacenan tripletas del tipo "sujeto a condiciones". Trabajando de esta manera, se espera asegurar que aquellas historias que aparecen con baja frecuencia, también sean tomadas en cuenta a la hora de hacer reconocimiento y generación de oraciones. Está claro que al dejar por fuera muchas historias de tres palabras, el rendimiento del sistema disminuye, sin embargo, pensamos que tenemos varias alternativas: una sería codificar y almacenar cada una de las historias que aparecen en el corpus, otra sería agregar más oraciones al corpus donde aparezcan con más frecuencia las historias antes rechazadas, es decir re-entrenando el modelo.

La segunda alternativa es la que hemos usado en las pruebas, puesto que el modelo que se crea de esta manera es adaptativo debido a que se puede repetir el procedimiento con otro conjunto de entrenamiento del contexto, y re-ajustar el modelo sin perder la información obtenida en un ajuste previo. Esto hace que en forma incremental se pueda enriquecer no sólo el vocabulario del modelo, sino que se puede aumentar también su capacidad para generar oraciones y para hacer la descodificación linguística.

El proceso de re-entrenamiento o re-ajuste consiste en que una vez que se tiene un modelo creado a partir de un corpus inicial al que llamaremos Corpus1, se puede seleccionar otras oraciones del mismo contexto, pero diferentes a las de dicho corpus, y elaborar con éstas un segundo corpus, digamos el Corpus2. Utilizando este nuevo corpus se codifican todas sus historias y se suman las frecuencias de aparición de aquellas que se encuentran tanto en el Corpus1 como en el Corpus2, por lo que se tendría una frecuencia acumulada para esas historias. Finalmente se almacena en los bloques mencionados 1, 2, 3 y 4, las nuevas palabras que inician oraciones, las historias pares de mayor frecuencia, las historias triples que contienen como sus dos palabras iniciales aquellas que constituyen las historias pares eliminadas y las nuevas tripletas que finalizan oraciones. Podemos ver que se trata de una

extensión del modelo creado con el Corpus1, al que se agregan nuevas historias y hasta nuevas palabras a su vocabulario, puesto que en el segundo corpus, pueden aparecer no sólo historias que no estaban presentes en el primero, sino también nuevas palabras. Este proceso se puede repetir con tantos corpus como se considere necesario, hasta obtener resultados adecuados en la generación de oraciones y por lo tanto en la descodificación lingüística.

11.6. GENERADOR DE ORACIONES. ALGORITMO PROPUESTO.

A partir del modelo codificado como se explicó en la sección anterior, el sistema es capaz de producir en forma aleatoria oraciones que están dentro del contexto que se modela.

El algoritmo para generar una oración es el siguiente:

1.- Se escoge en forma aleatoria una palabra de la lista del bloque1, digamos w_1, y se muestra en la pantalla del sistema. Por ejemplo, se escoge del bloque1, la palabra "Esta".

2.- Del bloque3, se agrupan aquellas historias que se inician con la palabra w_1 en un bloque nuevo, el sub-bloque31. Por ejemplo, "Esta especificación debe", "Esta oración declarativa", "Esta oración corresponde", "Esta distinción permite", "Esta juega un", etc.

3.- Si el sub-bloque31 resulta no vacío, se escoge aleatoriamente una de las historias que contiene y se muestran en la pantalla las dos palabras que siguen a w_1 en esa historia. Por ejemplo, se escoge la historia "Esta oración corresponde", y se muestra en pantalla las palabras "oración corresponde".

4.- Se continúa la búsqueda de historias en el bloque2. Las historias que interesan de este bloque, son aquellas que tienen como palabra inicial, la última que aparece en la historia seleccionada del sub-bloque31, cuando éste bloque es no vacío, digamos historias que comiencen con w_3. Cuando el sub-bloque31 es vacío, las historias que interesan son aquellas que tienen como palabra inicial a w_1. Se forma así un nuevo bloque, el sub-bloque21.

Supongamos que para este ejemplo, "corresponde a", sea la única historia presente en el sub-bloque21.

5.- Si el sub-bloque21 es no vacío, se selecciona aleatoriamente una de sus historias y se muestra la segunda palabra de esa historia.

En este ejemplo, se escoge la historia "corresponde a" porque es la única, pero en caso de haber más de una, se selecciona una de ellas aleatoriamente. Se muestra la segunda palabra, "a".

Hasta este momento se ha construido la frase "ESTA ORACION CORRESPONDE A".

6.- Se continua la búsqueda en el bloque4. Las historias que interesan de este bloque, son aquellas que comienzan con la última palabra de la última historia seleccionada en los pasos previos. Se forma el sub-bloque41.

Por ejemplo, las historias que interesan de este bloque serían las tripletas que comienzan con "a". En este ejemplo, supongamos que no se encuentran en el bloque4 historias que comiencen con "a". Es decir que el sub-bloque41 es vacío.

7.- Si el sub-bloque41 es no vacío, entonces se selecciona aleatoriamente una historia y se muestran sus dos últimas palabras. Aquí finalizaría la generación de una oración.

8.- Se actualiza w_1 con el código de la palabra con la que finaliza la última historia seleccionada.

Para el ejemplo, que se presenta aquí, w_1 tomaría el índice de "a".

9.- Se vuelve al paso 2.

Para el ejemplo descrito, después de volver al paso 2 y encontrar una historia en el bloque4, se obtiene la oración "ESTA ORACION CORRESPONDE A UN RITMO SILABICO".

La mayoría de las oraciones que se generan de esta manera no aparecen en el (los) corpus de entrenamiento, es decir, se forman a través de la conexión adecuada de las historias codificadas.

La oración que se generó en el ejemplo no aparece en el corpus, pero si aparecen las oraciones que se muestran a continuación y que por sus componentes se puede observar que intervienen mucho en la producción de la oración mencionada.

a.- ESTE TIPO DE ORACION CORRESPONDE A UN ENUNCIADO NEUTRO DESPROVISTO DE ASPECTOS EXPRESIVOS Y APELATIVOS ESPECIALES.

b.- CORRESPONDE A LA ULTIMA SILABA PORTADORA DE ACENTO LEXICO EN EL GRUPO MELODICO.

c.- ESTA ORACION DECLARATIVA ESTA CONSTITUIDA POR TRES UNIDADES TONALES.

En resumen, la generación de una oración, se hace entonces con una búsqueda sucesiva de palabras en las historias de los bloque3, bloque2 y bloque4, partiendo de la escogencia aleatoria de un índice del bloque1.

El proceso finaliza cuando se encuentra al menos una historia en el bloque final (bloque4) o cuando no aparece ninguna historia en ninguno de los tres bloques, que pueda continuar a una precedente. Este caso se puede presentar, cuando se selecciona una historia que en el corpus está ubicada al final de una oración; estas historias con frecuencia finalizan con palabras que no forman otras historias, por lo tanto ninguna historia las podrá seguir.

En este trabajo, se diseñó y construyó (por programación) el generador de oraciones descrito, con el fin de tener una idea de las secuencias de palabras que podría reconocer el descodificador lingüístico, que posteriormente se desarrolló y que se describe en la próxima sección. En este momento, podemos suponer que el modelo de contexto es una red de historias codificadas, que contiene las oraciones que pueden ser reconocidas por el descodificador

linguístico. La utilidad del generador de oraciones en este trabajo fue prevista, sólo para mostrar las oraciones presentes en el modelo del lenguaje de aplicaciones de reconocedores y que por lo tanto pueden ser reconocidas.

11.7. RECONOCEDOR O DESCODIFICADOR LINGÜÍSTICO. ALGORITMO PROPUESTO.

Recordemos que la parte de los reconocedores que convierte los datos acústicos de una pronunciación en una secuencia de símbolos linguísticos (por ejemplo, una secuencia de fonos, una secuencia de palabras, etc.) se llama Descodificador Acústico, mientras que el Descodificador Linguístico es la parte que determina si esa secuencia de símbolos, corresponde a una oración válida del lenguaje de la aplicación.

Tal como se aprecia en la figura 11.1, en este trabajo sólo se desarrolla el descodificador lingüístico, por lo que sus pruebas se realizan suponiendo que de existir un descodificador acústico, recibiría de éste una secuencia de palabras.

El procedimiento a través del cual el sistema puede reconocer una secuencia de palabras (w_1, w_2, ..., w_n), como gramaticalmente correcta a partir del modelo de contexto, se presenta a continuación:

1.- Recibe la primera palabra de la secuencia, w_1.

2.- Averigua si w_1 está presente en el vocabulario. Si w_1 no forma parte del vocabulario, la rechaza y por lo tanto, a la secuencia por estar fuera del contexto. Si pertenece al vocabulario muestra w_1 en la pantalla.

3.-Busca historias que se inicien con w_1 en bloque2 y bloque3. De esta manera se generan dos nuevos bloques de posibles partes de oraciones, que de acuerdo al lenguaje o contexto que modelan podrían formarse partiendo de w_1. Uno de esos bloques es producto de la búsqueda en bloque2, llamémoslo bloque21 y otro, producto de la búsqueda en bloque3, el bloque31. Se generan de esta manera, hipótesis parciales de oraciones. Esto constituye, creemos, una forma

para agilizar el proceso de reconocimiento, puesto que la búsqueda de la palabra siguiente, w_2, de la secuencia a reconocer se haría solo en bloque21 y bloque31.

Puede ocurrir que no se encuentren historias que se inicien con w_1, es decir, puede ocurrir que el bloque21 y el bloque31 resulten vacíos. En este caso, en el modelo no hay palabras que puedan seguir a w_1, por lo que el reconocedor no admitirá la oración (la secuencia de palabras) y finalizará el reconocimiento.

4.- Se recibe la siguiente palabra de la secuencia, w_2. Si forma parte del vocabulario, entonces se busca su ocurrencia en las historias presentes en el bloque21 y en el bloque31, en caso contrario se rechaza la secuencia.

5.- Si el bloque21 ó el bloque31 son no vacíos, se descartan de estos bloques aquellas historias que no contengan a w_2 después de w_1. Si quedan historias que contengan a w_2 siguiendo a w_1 se muestra w_2 en pantalla, en caso contrario se rechaza la secuencia y se termina su reconocimiento. Esto puede ocurrir, cuando el par (w_1 w_2) no aparece en el corpus de entrenamiento.

6.- Se vuelve al punto 3, trabajando con w_2 en lugar de w_1, es decir, cada vez que la descodificación llega a este punto, se re-inicia el recorrido trabajando con la última palabra tratada en el recorrido previo.

El reconocimiento de la secuencia de palabras tiene dos formas de finalización: una, cuando el descodificador la rechaza debido a que según el modelo no es válida o porque no pertenece al contexto y otra, cuando se recibe el símbolo $, que es el indicador de fin de oración (se escogió este símbolo debido a que en nuestro lenguaje escrito es poco frecuente su uso), en este último caso se tendrá una oración gramaticalmente correcta de la aplicación.

Por ejemplo, el descodificador linguístico reconocería como válida la secuencia "esta oración corresponde a un rítmo silábico $" (suponiendo que dicha secuencia la recibe desde un descodificador acústico cuya salida sean palabras), puesto que hemos visto que el generador puede producir dicha oración.

Se puede dar el caso de que se rechacen secuencias que pertenezcan al contexto y que sean gramaticalmente válidas. Esto se puede superar re-entrenando el modelo con nuevos corpus del mismo contexto.

Se puede apreciar que el reconocedor revisa, sí la secuencia de palabras que recibe es correcta desde el punto de vista de las reglas gramaticales del lenguaje al cual está asociado el contexto y determina también, sí forma parte del contexto que se modela.

Aunque los modelos de contextos descritos pueden pensarse como una combinación de Bigramas y Trigramas [22][33][56], en este trabajo no podemos hablar de modelos n-gramas estocásticos, ni de autómatas de estados finitos estocásticos [1][22][33], puesto que para la forma como se hace la descodificación no se utilizan las probabilidades. De hecho, este descodificador no mide la probabilidad de que las secuencias estén modeladas o no, simplemente, sí puede formar una oración que esté en el modelo la acepta, de lo contrario la rechaza.

11.8. PRUEBAS DEL SISTEMA

Los ensayos que se hicieron consistieron en:

1.- Se realizó una prueba inicial con un corpus formado por 160 oraciones y párrafos de distintas longitudes. Se trabajó con longitudes de entre tres y sesenta y tres palabras. Las oraciones fueron tomadas de un texto propio de la lingüística. El corpus completo comprendía 2563 palabras.

2.- Se obtuvo el vocabulario que manejarían tanto el módulo reconocedor como el módulo generador. El vocabulario al principio fue de 816 palabras.

3.- Se formaron los bloques: bloque1, bloque2, bloque3 y el bloque4 que constituyeron el modelo del contexto.

4.- Se generaron bloques de oraciones. Este proceso se repitió unas 30 veces. Cada bloque generado, comprendía diez oraciones.

5.- Se realizó el reconocimiento de oraciones. La prueba consistió en que dadas secuencias de palabras, se averiguaba si dicha secuencias, podían ser reconocidas usando el modelo del contexto.

6.- Se tomaron de nuevo, pequeños corpus de 10 y 20 oraciones y se repitió el procedimiento.

11.9. RESULTADOS DE LAS PRUEBAS

1.- Se pudo utilizar nuevos corpus en forma incremental, sin que se perdiera la información codificada en ensayos anteriores.

2.- Las oraciones generadas, eran en general, más pequeñas en longitud respecto a las contenidas en el corpus de entrenamiento.

3.- El número de oraciones generadas dependía del tamaño del corpus de entrenamiento.

4.- El número de oraciones generadas que eran válidas, en cuanto a la gramática y al contexto era aproximadamente el 70% del total de las que se generaron en los ensayos.

5.- El número de oraciones reconocidas que eran gramaticalmente correctas y que pertenecían al contexto era cercano al 90%, cuando esas oraciones se escogieron muy parecidas a las del corpus de entrenamiento.

6.- Aproximadamente el 90% de las oraciones generadas no pertenecían al corpus de entrenamiento, a excepción de algunas oraciones cortas, de tres, cuatro y hasta cinco palabras.

7.- No era posible reconocer todas las oraciones y párrafos, tal como aparecían en el corpus de entrenamiento.

11.10. CONCLUSIONES

En forma incremental se puede lograr mejorar la robustez tanto del módulo reconocedor como del generador de oraciones. Claro está, esto conlleva lentitud durante el re-ajuste del modelo, que dependerá de la aplicación y de la máquina.

Como se pueden generar grandes cantidades de oraciones, se puede también reconocer un número grande de frases y oraciones.

Debido a la gran cantidad de oraciones que se pueden generar y que no pertenecen al corpus de entrenamiento, es posible, también reconocer una gran cantidad de oraciones y frases no necesariamente propias del contexto que se modela, pero si propias del lenguaje en el que está escrito el corpus.

No es posible reconocer todas las oraciones y párrafos, tal como aparecen en el corpus de entrenamiento, debido a que en la memoria del sistema no aparecen todas las historias presentes en el corpus. Lo que podría superarse si se almacenan todas las historias que aparecen en el texto, pero esto conllevaría a que la búsqueda tanto en reconocimiento como en generación sea más lenta.

Trabajando con la codificación y los tamaños de las historias que se han indicado, se puede crear modelos aceptables de contextos.

Se trata de un reconocedor, un generador de oraciones y un modelador de lenguaje de aplicaciones, altamente determinístico.

Este tipo de descodificador lingüístico, podría funcionar en aplicaciones de reconocimiento donde el tamaño del vocabulario abarque varios miles de palabras.

CAPITULO XII

CONCLUSIONES, CONTRIBUCIONES Y RECOMENDACIONES GENERALES.

Los experimentos realizados durante la ejecución de esta tesis han dado lugar a las conclusiones siguientes:

- Con modelos de unas pocas palabras, construidos a través de la voz de hombres y de mujeres de Venezuela, por separado, es posible que un reconocedor pueda determinar el género del hablante. Una estrategia válida, sería pedirle a cada locutor que pronuncie en secuencia varias de las palabras modeladas y que el reconocedor obtenga el porcentaje de las palabras que correspondan a hombres y el porcentaje de las palabras que correspondan a mujeres; el porcentaje mayor determinará el género del locutor. Esta fue la manera en que se obtuvieron buenos resultados, cuando se crearon los modelos de los dígitos pronunciados por hombres y los modelos de los dígitos pronunciados por mujeres.

- Para construir reconocedores para aplicaciones donde las entradas sean cadenas de dígitos pronunciados por venezolanos, se puede recurrir a modelos de palabras (uno por dígito) y a modelos de fonos, puesto que el nivel de reconocimiento en los dos tipos de pruebas resultó suficientemente alto. Claro está, siempre habrá que recurrir a un mecanismo que asegure que el porcentaje del error en el reconocimiento sea mínimo, para ello podemos indicar que una forma de minimizar el error, sería solicitándole al hablante que diga la secuencia varias veces, y que se acepte como reconocida, aquella secuencia de salida donde los dígitos presenten mayor frecuencia de aparición.

- Es posible crear reconocedores de la voz venezolana, independientes del hablante.

- El reconocimiento automático de los dialectos venezolanos, es imposible realizarlo a través de modelos de palabras, cuando esas palabras provienen de texto leído.

- De los resultados de las pruebas de reconocimiento de dialectos venezolanos, donde se trabajó con modelos de fonos, hay indicios de que se puede alcanzar un nivel aceptable de reconocimiento.

- Para alcanzar un porcentaje alto, cuando se hace reconocimiento de oraciones, hay que elaborar de manera muy fina la gramática del lenguaje de la aplicación, para minimizar los errores en los que incurre el reconocedor de las unidades acústicas. Esto se pudo observar cuando se realizó reconocimiento de fechas y modelado de lenguajes de aplicaciones.

- Para cada tipo de aplicación, hay que construir modelos de la voz propia de esa aplicación, para alcanzar mejor rendimiento del sistema de reconocimiento. Esto se pudo observar, cuando se realizaron pruebas de reconocimiento de cadenas de dígitos, donde se partía de los modelos de fonos construidos con las pronunciaciones de fechas; el nivel de reconocimiento era muy bajo respecto al resultado que se obtuvo cuando se trabajó con modelos construidos a partir de pronunciaciones de dígitos solamente.

- Es posible construir modelos de fonos, fonemas y palabras de la voz venezolana para hacer reconocimiento de palabras aisladas, de palabras conectadas y de habla contínua.

- Para muchas aplicaciones de reconocimiento, puede ser conveniente crear modelos de la voz de mujeres por separado de los modelos de la voz de hombres, puesto que modelos globales (que representen las características de locutores de ambos sexos) robustos, son más difíciles de obtener.

Consideramos que los logros más importantes de este trabajo son los siguientes:

- Introducir este tipo de Tecnología en Venezuela.

- Introducir la teoría de los Modelos Ocultos de Markov, para modelado del habla venezolana, en la Universidad de Los Andes en Mérida-Venezuela.

- Realizar en Venezuela, los primeros trabajos de modelado matemático/acústico de fonemas y palabras del habla de los venezolanos para fines de reconocimiento automático.

- Realizar los primeros trabajos en Venezuela, donde se hace reconocimiento automático de habla venezolana.

- Realizar las primeras publicaciones sobre el reconocimiento automático del español hablado en Venezuela.

- Conformar conjuntamente con los profesores Manuel Rodríguez y Elsa Mora, una estructura básica para crear un grupo de trabajo en Tecnología del habla en la Universidad de Los Andes, Mérida-Venezuela.

- Producir el punto de partida para la construcción de sistemas de reconocimiento del habla venezolana para aplicaciones reales.

En cuanto a las recomendaciones que podríamos indicar para futuros trabajos, donde se trate la voz venezolana en forma automática tenemos que:

- Aun cuando, en la mayoría de las pruebas el nivel de reconocimiento alcanzó porcentajes suficientemente altos, muchas de estas pruebas se realizaron con una fracción de archivos de voz de la SPEECHDAT Venezolana, por lo que se recomienda, verificar qué niveles de reconocimiento se pueden alcanzar cuándo se utiliza una proporción mayor de dicha base de datos. Para ello, se pueden usar los mismos programas escritos para HTK en estas pruebas, sólo tendría que agregarse nuevas listas de datos.

- Se puede iniciar la construcción de sistemas de reconocimiento venezolanos para algunas aplicaciones particulares, donde se requiera como entrada el tipo de señales que se manejaron en este trabajo. Para ello se dispone de todos los modelos estimados/entrenados en estas pruebas.

- Debido a que contamos con una base de datos de la voz venezolana, se podrá realizar todo tipo de pruebas desde el punto de vista del reconocimiento automático y desde el punto de vista lingüístico.

- Se debe continuar con esta línea de trabajo.

BIBLIOGRAFIA

[1] Deller J., Proakis J. y Jansen J., Discrete-Time Processing of Speech Signals. New York : Macmilllan Publishing Company, 1993.

[2] Moreno A., Mora E., Speechdat Spanish Venezuelan database for the fixed Telephone network, Universidad politécnica de Cataluña, España y Universidad de Los Andes, Venezuela, 1999.

[3] Moreno A., Speechdat Spanish Database for the fixed Telephone Network. Universidad Politécnica de Cataluña, España. 1997.

[4] Entropic Speech Technology, The HTK Book, version 2.2, (1999).

[5] Savage J., A hybrid System with Symbolic AI and Statistical Methods for Speech Recognition. A dissertation for the degree of Doctor of Philosophy, University of Washington, USA 1995.

[6] Zhao Y., A Speaker-Independent Continuous Speech Recognition System using Continuous mixture Gaussian Density HMM of phoneme-sized units. IEEE Transaction on Speech and Audio Processing, Vol. 1, No. 3, July (1993).

[7] Furui S., Digital Speech Processing, Synthesis and Recognition, New York; Marcel Dekker, 1989.

[8] Rabiner L. and Juang B., Fundamentals of Speech Recognition, Englewood Cliffs,New Jersey; Prentice Hall, 1993.

[9] Picone J., Signal Modeling Techniques in Speech Recognition, Proc. IEEE, 81(9), 1215-1246, september, 1993.

[10] Neukirchen C., "Exploiting Acoustic Feature Correlations By Joint Neural vector Quantizer Design in a Discrete HMM System". International Conference on Acoustics, Speech and Signal Processing. Seattle, USA 1998.

[11] Chengalvarayan R., "On the use of normalized LPC error towards better large vocabulary speech recognition systems". International Conference on Acoustics, Speech and Signal Processing. Seattle, USA 1998. Speech Processing Group, Bell Labs.

[12] Casacuberta F. y Vidal E., Reconocimiento Automático del Habla. Marcombo, Boixareu Editores, Barcelona-Mexico, 1987.

[13] Paul D., "Speech Recognition Using Hidden Markov Models". The Lincoln Laboratory Journal, Volume 3, Number 1, 1990.

[14] Juang B., Rabiner L., "Issues in Using Hidden Markov Models for Speech Recognition". Speech Research Department, AT&T Bell Laboratories, 1992.

[15] Picone J., "Continuous Speech Recognition Using Hidden Markov Models". IEEE ASSP Magazine, pp. 26-41. July 1990.

[16] Torres I. and Casacuberta F., "Spanish Phone Recognition Using Semicontinuos Hidden Models". Universidad del País Vasco y Universidad Politécnica de Valencia, España.

[17] Rabiner L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of The IEEE, Vol. 77, NO. 2, February 1989.

[18] Dayhoff J., Neural Networks Arquitectures. Editorial Van Nostrand Reinhold, New York, 1990.

[19] Waibel A. and Hampshire J., "Building Blocks for Speech". BYTE, Agust 1989.

[20] Simon H., Neural Networks, A comprehensive Foundation. MAC MILLAN, IEEE PRESS, 1999.

[21] Mariño J., "Generation of multiple hypothesis in connected phonetic-unit recognition by a modified one-stage dynamic programming algorithm". Universidad Politécnica de Cataluña, España. EUROSPEECH 1989.

[22] Bonafonte A., Comprensión del Habla en Tareas Semánticamente Restringidas. Tesis Doctoral. Universidad Politécnica de Cataluña, Barcelona, 1995.

[23] Liu W., The NTN/HMM Hybrid Keyword Spotter. Master´s Thesis. Rutgers - The state University of New Jersey, USA, 1996.

[24] Savage-Carmona, A Hybrid System with Symbolic AI and Statistical Methods for Speech Recognition. A dissertation for the degree of Doctor of Philosophy. University of Washington, USA, 1995.

[25] Kanungo T., Hidden Markov Models. University of Maryland, USA, 1998.

[26] Bonafonte A. y Moreno A., "Reconocimiento de palabras aisladas mediante Modelos Ocultos de Markov". ETS. D'Enginyería de Telecomunicació, Barcelona, Marzo 1998.

[27] Lippmann R., "An Introduction to computing with Neural Nets". IEEE ASSP Magazine April 1987.

[28] Alvarez A., Martínez R., Nieto V., Rodellar V. y Gómez P., A Robust Isolated Word Recognizer for Higly Non-Stationary Environments. Recognition Results, EUROSPEECH`99, Budapest, Hungary, vol. 6, 2817-2820, September (1999).

[29] Cumana J., Reconocimiento Automático de Voz: Dígitos continuos con modelos Markovianos Ocultos, Universidad de Los Andes, Mérida-Venezuela, 2001.

[30] Gish H. and Schmidt M., Text-Independent Speaker Identification, IEEE Signal Processing Magazine October (1994).

[31] Itahashi S. and Du L., Language Identification based on Speech Fundamental Frecuency, EUROSPEECH`95, 1359-1361, (1995).

[32] Hoge H., Moreno A., Technical Annex of the SALA Contract. SALA project feb. 1999.

[33] Bonafonte A. and Mariño J., "Language Modeling using X-Grams", International Conference on Spoken Language Processing, ICSLP-96.

[34] Maldonado J. L, El Estado del Arte de la Tecnología del Habla. Charla dictada en el ciclo de Seminarios del postgrado en Ingeniería de Control, Facultad de Ingeniería, ULA. Julio de 1997.

[35] Maldonado J. L, Algorítmos para el Reconocimiento Automático del Habla. Charla dictada en el ciclo de Seminarios del postgrado en Ingeniería de Control, Facultad de Ingeniería, ULA. Marzo de 1998.

[36] Maldonado J. L., Una aplicación de las Redes Neurales Artificiales al Reconocimiento de las Vocales expresadas en Español. Tesis de Grado para optar al título de Magister Scientiae en Ingeniería de Control. Facultad de Ingeniería, ULA, 1994.

[37] Neural Networks. Electronics World + Wireless World.August 1993.

[38] Mora E., Caractérisation prosodique de la variation dialectale de l'Espagnol parlé au Vénézuéla. Thesis de Doctorat en Phonetique experimentale, fonctionnelle et appliquee. Universite de Provence-Aix-Marseille I, la France, 1996.

[39] Huang X., Acero A., Hon H., Spoken Language Processing, A guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001.

[40] Chomsky N., Aspectos de la teoría de la Sintaxis.
Gedisa editorial, 1999

[41] Obediente E., Fonética y Fonología.
Universidad de Los Andes, Venezuela, 1991.

[42] Ontoso R., Entrenamiento de árboles de decisión para la generalización de unidades fonéticas contextuales en el reconocimiento del habla contínua.
Grupo de tratamiento del Habla. Proyecto de grado, UPC, Barcelona, España, 1999.

[43] Rodríguez M., Procesamiento digital de señales.
Universidad de Los Andes, Venezuela.

[44] James A., Natural Language Unserstanding.
The Benjamin/Cummings Publishing Company, Inc., 1989.

[45] Pachés P., Improved modelling for robust Speech Recognition.
Grupo de tratamiento del Habla. Tesis doctoral, UPC, Barcelona, España, 1999.

[46] Información sobre el formato Speechdat para construcción de corpus orales contenida en:
www.speechdat.org/speechdat.html

[47] Información sobre el formato Speechdat para construcción de corpus orales contenida en:
www.speechdat.org/speechdat.html/wwwTranscribe.pdf

[48] De la torre A., Reconocimiento Automático de Voz en Condiciones de Ruido.
Tesis doctoral, Universidad de Granada, España, 2001.

[49] Mora E., División Prosódica Dialectal de Venezuela.
OMNIA, Revista interdisciplinaria de la División para graduados de la Facultada de Humanidades y Educación, Nro. 2, diciembre 1997, LUZ - Venezuela.

[50] Benítez M., Reconocimiento de Palabras Claves en Sistemas Independientes de la Tarea.
Tesis doctoral, Universidad de Granada, España, 1998.

[51] SPEX, Speech Processing EXpertise centre.
http://www.spex.nl/

[52] SAMPA ,Speech Assessment Methods Phonetic Alphabet.
www.phon.ucl.ac.uk/home/sampa/home.htm

[53] VoxStudio software.
http://www.xentec.be/

[54] Telephony Based Speaker-Independent Large Vocabulary Continuous Mandarin Speech Recognition.
http://wwwckip.iis.sinica.edu.tw/CLCLP/Vol4-1/a1.htm

[55] Modeling Structure in Speech above the Segment for Spontaneous Speech Recognition.
http://cslu.cse.ogi.edu/nsf/isgw97/reports/ostendorf1.html

[56] Cardenal A., Realización de un Reconocedor de Voz en Tiempo Real para Habla contínua y Grandes Vocabularios. Tesis doctoral, Universidad de Vigo, España, 2001.

[57] García P., Reconocimiento Automático de Voz contínua con Modelos Ocultos de Markov. Tesis doctoral, Universidad de Granada, España, 2001.

[58] Peinado A., Reconocimiento de Voz mediante Modelos Ocultos de Markov: Selección y Estimación de parámetros. Tesis doctoral, Universidad de Granada, España, 1994.

[59] Díaz J., Reconocimiento de Voz Contínua. Aproximaciones basadas en HMM y en Redes Neuronales Recurrentes. Tesis doctoral, Universidad de Granada, España, 2002.

[60] Maldonado J. L., Redes Neuronales Artificiales.
Curso dictado en la Universidad de Granada, España, marzo 2003.

[61] González B.y García M., Diseño de una Base de Datos tipo SpeechDat para el idioma Gallego. SEPLN, Sociedad española para el Procesamiento del lenguaje Natural, Revista Nro. 24, septiembre del 2000.

[62] Trabajo titulado "Application of *VQ* to *Q-learning*. VQQL" que se encuentra en la dirección electrónica, http://grial.uc3m.es/~ffernand/publicaciones/robocup99/node6.html

[63] Hilera J., y Martínez V., Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones. Alfaomega RA-MA Editorial, Madrid, España, 2000.

[64] Hagan M. y Demuth H., Neural Network Design.
PWS Publishing Company, 1999.

[65] Anderson L. y Jensen C., Paradox para Windows.
McGraw-Hill/Interamericana de España, S.A., 1995.

[66] Sánchez C., Microsoft Access 2000.
McGraw-Hill, España, 1999.

[67] Mora E., Producción del Habla.
Monografía, ULA, 2002.

[68] Applied Technologies on Language and Speech, ATLAS.
http://www.atlas-cti.com/

[69] Jelinek F., "Continuous Speech recognition by Statistical Methods".
IEEE Proceedings , Vol. 64, NO. 4, 1976.

[70] Jelinek F. et al., "A Maximum Likelihood Approach to Continuous Speech Recognition".
IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 5, NO. 2, 1983.

GLOSARIO DE TÉRMINOS

ANN: Redes Neurales Artificiales.

ALÓFONOS: Las variaciones acústicas de cada fonema.

A/D: Convertidores de señales analógicas a digitales.

BKP: Algoritmo de retropropagación del error (backpropagation algorithm).

CEPSTRUM: El conjunto de los parámetros obtenidos del análisis cepstral.

CLUSTERING: Proceso mediante el cual se crean grupos de vectores en términos de una medida de distancia.

COARTICULACIÓN: La transición entre los sonidos que constituyen una pronunciación.

DTFT: Transformada de Fourier en tiempo discreto.

FONO: La realización concreta de un fonema, es decir, un sonido del habla (la realización acústica de un fonema).

FONEMA: Unidad básica a través de la cual se describen teóricamente los sonidos de la lengua.

FORMANTES: Las resonancias del espectro de magnitud de la voz.

FRECUENCIA FUNDAMENTAL: Definida como el inverso del período entre aperturas sucesivas de las cuerdas vocales (campo de reconocimiento).

GRAFEMAS: Escritura ordinaria de las palabras.

HTK: Hidden Markov Models Toolkit, herramienta de programación diseñado para construir reconocedores del habla basados en MOM.

IDTFT: Transformada inversa de Fourier en tiempo discreto.

LPC: Códigos de Predicción Lineal.

LP: Predicción Lineal.

LBG: Algoritmo de Cuantificación vectorial de Linde, Buzo y Gray.

LOCUTOR: Persona cuya voz se utiliza para hacer entrenamiento o reconocimiento automático; en esta tesis se usa indistintamente del término hablante.

LVQ: El algoritmo Learning Vector Quantizer de Kohonen.

MFCC: Coeficientes cepstrales en escala de frecuencia Mel.

MOM: Modelos Ocultos de Markov.

MU-LAW: La codificación estándar Norteamericana, U-Law.

MSE: Error cuadrático medio.

PITCH: Es la frecuencia fundamental de la vibración de las cuerdas vocales.

PO: Principio de Ortogonalidad.

RECONOCEDORES: En este contexto nos referimos a sistemas de reconocimiento automáticos de la voz.

SAMPA: Speech Assessment Methods Phonetic Alphabet.

SAGA: Spanish automatic graphemes to allophones transcriber, propiedad de la UPC.

SONIDOS SONOROS: sonidos de voz producidos con alta periodicidad de la señal de excitación producida por la glotis.

SONIDOS SORDOS: sonidos de voz producidos con ninguna periodicidad de la señal de excitación producida por la presión de aire que parte de los pulmones.

UPC: Universidad Politécnica de Cataluña, España.

VOCABULARIO: Conjunto de palabras diferentes que pueden presentar como salida, los sistemas de reconocimiento.

ANEXOS

ANEXO A

LOS MODELOS OCULTOS DE MARKOV

A.1. INTRODUCCIÓN

La razón por la cual se dedica este espacio a esas técnicas se debe a que constituyen la base de la mayoría de las pruebas de modelado de voz realizadas en este trabajo, y a que son las más utilizadas en campo del reconocimiento automático del habla.

A.2. ORIGEN DE LOS MODELOS OCULTOS DE MARKOV

La historia de los MOM se remonta a los años cincuenta del siglo pasado, cuando estudiosos de la Estadística estaban tratando el problema de caracterizar procesos aleatorios para los cuales no se contaba con suficientes observaciones [1].

La idea implicaba modelar el problema como un proceso estocástico doble, en el cual los datos observados (las realizaciones) se asumían como el resultado de generar realizaciones de un primer proceso aleatorio (el proceso oculto), a través de un medio que daba origen a un segundo proceso aleatorio (el proceso observado). Los dos procesos se lograban caracterizar usando sólo el que se podía observar [1][50][57][58].

Del estudio de este problema surgió el algorítmo de identificación que se conoce como el **algorítmo de máxima estimación (ME)** [1]. Luego, en los primeros años de la década de mil novecientos setenta se desarrolló un caso especial del algorítmo de ME, el F-B (forward-backward) también llamado algorítmo de re-estimación Baum-Welch, en honor a sus generadores, para estimar los parámetros de los MOM, tal como se usa en la actualidad en el problema del reconocimiento de la voz y en otras aplicaciones.

A.3. LA IDEA DETRÁS DE LOS MOM

Los sistemas del mundo real en general producen salidas o datos que se pueden tratar como señales. Dichas señales pueden ser de naturaleza discreta (por ejemplo, las salidas del lanzamiento sucesivo de un dado) o de naturaleza continua (por ejemplo, las medidas de la corriente eléctrica en un determinado ambiente). Estas señales pueden ser estacionarias o no estacionarias según varíen o no sus propiedades estadísticas a través del tiempo y pueden estar corrompidas o no por otras señales de su entorno [1][13][14][17].

En ese orden de ideas, existe el problema fundamental de crear modelos para esas señales con la finalidad de que a partir de éstos, se puedan describir teóricamente, simular, controlar y hasta construir los procesos generadores de dichas señales.

Los modelos de señales se clasifican en dos categorías: determinísticos y estocásticos. Los determinísticos en general, explotan propiedades conocidas de las señales (por ejemplo, si son sinusoides, si son sumas de exponenciales etc., tal vez sería suficiente seleccionar como valores de los parámetros del modelo, la amplitud, la frecuencia, el número de cruces por cero, etc.), mientras que en los estocásticos se intenta modelar solamente las propiedades estadísticas de la señal. Los Modelos Ocultos de Markov caen en esta última categoría.

Para fijar ideas respecto a los MOM, se puede revisar el experimento siguiente [17]: Considérese un sistema que consiste en que en algún lugar de un salón, hay un conjunto de cajas y bolas. Y que en el salón hay un grupo de personas; las personas se encuentran separadas del conjunto de las cajas y bolas, por una cortina, es decir, las personas no ven esos objetos. Supóngase que hay un número N de cajas numeradas y que cada caja contiene un número considerable de bolas de colores. Supóngase también, que hay K distintos colores para las bolas.

El proceso físico para obtener las observaciones es como sigue: una persona, llamémosla el Mago, entra al lugar donde están los objetos mencionados y escoge una caja, la caja inicial, de acuerdo a algún proceso aleatorio. De esta caja toma una bola al azar y la muestra por encima de la cortina al grupo de personas; las personas anotan el color observado. El Mago vuelve a colocar la bola en la caja de donde fue extraída. Luego, selecciona una nueva caja de acuerdo

al mismo proceso aleatorio con que seleccionó la primera, y toma una bola tal como lo hizo antes, y la muestra al público. Las personas vuelven a anotar el color.

Si se repite el procedimiento varias veces se tendrá una secuencia finita de observaciones de colores, que constituyen una realización del proceso observado. Aquí, se puede notar que hay un proceso aleatorio oculto que da origen a ese proceso observado; ese proceso aleatorio oculto es el proceso aleatorio de la secuencia de cajas, es decir, la gente no sabe de que caja o cajas provienen las bolas de colores, sin embargo, se dan cuenta que ocurre una realización al estilo caja 1, caja 3, caja 6, caja 7, caja 1, ..., caja 4, que no pueden observar, pero que da origen a la realización observada del tipo azul, rojo, azul, verde, verde,...., rojo.

Un MOM permite modelar este tipo de experimento partiendo de la secuencia de salidas observadas.

A.4. COMPONENTES DE LOS MODELOS OCULTOS DE MARKOV

Un MOM consiste formalmente de los siguientes elementos:

- El número de estados del modelo, **N.**

En los MOM los estados están ocultos en general (son difíciles de definir), sin embargo, para algunas aplicaciones prácticas tienen significado físico. En el experimento mencionado arriba, los estados corresponden a las cajas.

- El conjunto de los estados, **E = {1,2,...,N}.**

La secuencia de estados, es el primero de dos procesos aleatorios asociados con un MOM.

- El proceso aleatorio de los estados, $\underline{x}$.

- Las variables aleatorias asociadas al proceso aleatorio de los estados, $\underline{x}(t)$.

- El número de símbolos distintos que pueden ser observados en los estados (los distintos colores de las bolas del experimento descrito), **K.**
- El conjunto de los símbolos distintos, **V = {1,2,...,K}.**
- La longitud de la secuencia observable, **T.**
- La probabilidad de ocurrencia del estado i al inicio del experimento, (t=1),

$\prod(t) = [p(\underline{x}(t) = i]$.

- La matriz de probabilidades de transición de estado, **A[a(i/j)].**
- La probabilidad de que ocurra el estado i en el instante t, dado que en t-1 ocurrió el estado j,

a(i/j) = P($\underline{x}(t) =$ **i/** $\underline{x}(t-1) = j$ **),** para $1 \leq i,j \leq N$ y t arbitrario.

Observación 1: Las filas de **A** suman 1 debido a que en cualquier instante t, ocurre una transición. Se asume que las a(i/j) son independientes del tiempo, es decir, no cambian durante el experimento.

La secuencia de las observaciones se modela también como un proceso estocástico $\underline{y}$ (el segundo proceso aleatorio), con variables aleatorias, $\underline{y}(t)$, independientes e idénticamente distribuidas. Se asume que en un estado **i,** en un instante t, se genera una observación.

- La secuencia observada, **y = {y(1), y(2),...,y(t),...,y(T)}.**
- El símbolo observado en el instante t, y(t).
- La matriz de probabilidades de las observaciones, **B[b(y(t)/i)].**
- La probabilidad de que ocurra el símbolo y(t) en el estado i, en el instante t

b(y(t)/i) = p($\underline{y}(t)$ **=y(t)/** $\underline{x}(t)$ **=i).**

Observación 2: Las probabilidades de las observaciones dependen del estado y son independientes de t. Las filas de la matriz **B** suman 1, debido a que siempre se genera una observación en el estado i en cualquier instante t.

En sentido formal un MOM comprende la siguiente estructura matemática:

$$m = \{E, \prod(1), A, B\} \qquad (A.1)$$

Bajo la condición de no preocuparnos por la historia de la secuencia de estados, esta secuencia aleatoria (de primer orden) es un **proceso marcoviano** [1][14]. Recordemos también que una **Cadena de Markov** es un proceso marcoviano donde las variables aleatorias toman valores discretos, enteros en general.

A.5. DEFINICION DE LOS MODELOS OCULTOS DE MARKOV

Un MOM es una máquina abstracta que modela un proceso estocástico, es un autómata de estados finitos, donde la ocurrencia de estos estados está asociada con una distribución de probabilidad; y donde las transiciones entre los estados están gobernadas por un conjunto de probabilidades llamadas probabilidades de transición de estados. Además, en un estado particular, una observación se genera también de acuerdo a una distribución de probabilidad. Los estados no son visibles en general, de ahí el nombre de MOM [1][8][25].

En las figuras A.1 y A.2 se muestran dos MOM con topologías distintas y típicas.

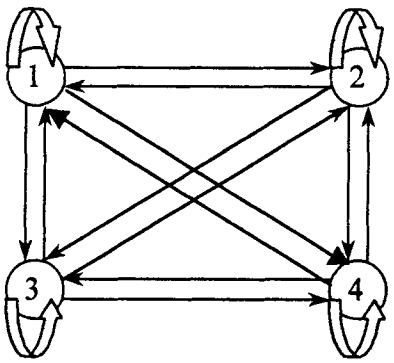


Figura A.1. Un MOM de cuatro estados completamente conectados

En la figura A.1. se puede ver un MOM de 4 estados que admite transiciones hacia cualquier otro estado, incluso hacia si mismo, en un instante cualquiera t.

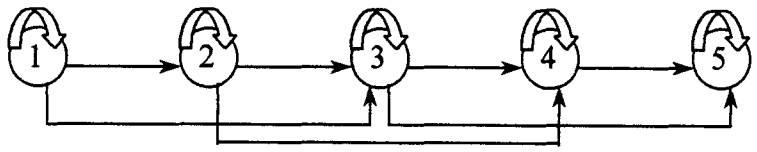


Figura A.2. Un MOM de 5 estados parcialmente conectados

En la figura A.2. se puede ver un MOM de 5 estados que admite transiciones a lo más, hacia los dos estados siguientes y a sí mismo, en un instante cualquiera t.

La selección de la topología obedece a la aplicación que se desee dar al MOM que se esté creando. En reconocimiento de voz se pueden usar diversas topologías, sin embargo, las que mejores resultados producen son las del tipo de izquierda a derecha al estilo de la figura A.2.

A.6. CLASIFICACION DE LOS MOM DE ACUERDO A LOS VALORES DE LAS OBSERVACIONES

Cuando el conjunto de símbolos K (distintos), con los que se forman las secuencias de observaciones, es muy grande (por encima de 256), se habla de secuencia de observaciones continuas. En este caso, se usa una función de densidad de probabilidades multivariante contínua, en lugar de un conjunto de probabilidades discretas para generar dichas observaciones [1]. En reconocimiento, se acostumbra usar para estos efectos una suma pesada de M distribuciones Gausianas G .

Lo anterior lleva a la clasificación de los MOM en: MOM de observaciones discretas y MOM de observaciones continuas.

A.7. LOS PROBLEMAS ASOCIADOS A LOS MOM

Una vez que se decide crear un MOM se encuentran tres problemas de interés por resolver [1][13][14][15][16][17][25].

1.- El problema de Evaluación: Dado un MOM **m** y una secuencia de observaciones y={y(1), y(2),...,y(t),...,y(T)}, cómo calcular la probabilidad de que esa secuencia sea generada por el modelo. Se tiene el problema de calcular de manera eficiente la probabilidad de que las observaciones sean generadas por el modelo, la probabilidad P(y/m).

2.- El problema de descodificación: Dado un MOM **m** y una secuencia de observaciones y={y(1), y(2),...,y(t),...,y(T)}, hay que determinar la mejor secuencia de estados en el modelo, que produce a esa secuencia de observaciones. Se tiene el problema de encontrar la secuencia de estados que haga máxima a P(y/m).

3.- El problema de entrenamiento: Dado un MOM **m** y una secuencia de observaciones y={y(1), y(2),...,y(t),...,y(T)}, cómo ajustar los parámetros del modelo. Se está frente al problema de calcular el mejor conjunto $\{E, \prod^{(l)}, A, B\}$ con el fin de maximizar P(y/m).

A.8. LOS MODELOS OCULTOS DE MARKOV DISCRETOS

Los MOM discretos son aquellos donde cada una de sus observaciones toma un valor entre un conjunto de **K** valores finitos (por debajo de 256 valores distintos), y donde en cada estado **i**, la función de distribución de probabilidad, que genera a esas observaciones, toma la forma de **K** impulsos en la recta real [1].

Un MOM de observaciones discretas se expresa como:

$$\mathbf{m}=\{E, \prod^{(l)}, A, B, \{y_k, 1 \le k \le K\}\} \qquad (A.2)$$

donde $\{y_k, 1 \le k \le K\}$ es el conjunto de los K símbolos distintos que puede generar el MOM en un estado.

A.9. SOLUCION DE LOS TRES PROBLEMAS ASOCIADOS A LOS MOM DE OBSERVACIONES DISCRETAS

Cuando se va a construir un Modelo Oculto de Markov, se parte de un modelo más o menos arbitrario, en el sentido de que la matriz de probabilidades de transición de estados y la matriz de probabilidades de producir las observaciones en un estado, contienen valores arbitrarios, pero por supuesto, deben ser valores entre cero y uno (valores probabilísticos), y además, deben cumplir con la regla de probabilidades, de que la suma de las probabilidades de transición desde un estado a si mismo y al resto debe dar uno; de la misma manera, la suma de las probabilidades de que un estado produzca, represente o modele cada observación debe dar igualmente uno.

En reconocimiento de voz, lo único que se sabe con cierta certeza es que cada modelo debe contener entre 2 y 5 estados, cuando se modelan unidades de palabras (fonemas, fonos, difonos, semifonemas, trifonemas, sílabas, etc.), y entre 5 y 12 estados cuando se modelan palabras [1][15][26][45]. La justificación del uso de ese número de estados por modelo está en que de esa manera es que se han logrado los mejores resultados.

A continuación se presentan diversas técnicas, a través de las cuales se superan los tres problemas relacionados con la construcción de los MOM.

A.9.1. SOLUCION AL PROBLEMA DE EVALUACIÓN DE LOS MOM DISCRETOS

Supóngase que se tiene una secuencia de observaciones de la forma $\mathbf{y} = y_1^T$ = {y(1),...,y(T)} y un modelo **m**, se presentan dos enfoques para calcular la probabilidad de que tal secuencia sea producida o pueda ser representada por el modelo **m**.

EL ENFOQUE FORWARD-BACKWARD, F-B:

Las observaciones se producen usando cualquier secuencia de estados en el MOM **m**.

Sea L una secuencia específica de estados de longitud T,

$L=\{i_1, i_2, \ldots, i_T\}$.

La probabilidad de que la secuencia de observaciones y_1^T, sea producida por la secuencia de estados L es:

$$\mathbf{P(y/L,m)} = \mathbf{b(y(1)/}i_1\mathbf{)b(y(2)/}i_2\mathbf{)...b(y(T)/}i_T\mathbf{)}, \tag{A.3}$$

y la probabilidad de que ocurra la secuencia L en el modelo **m** está dada por:

$$\mathbf{P(L/m)} = \mathbf{P(\underline{x}(1)=}i_1\mathbf{)a(}i_2/i_1\mathbf{)a(}i_3/i_2\mathbf{)....a(}i_T/i_{T-1}\mathbf{)}. \tag{A.4}$$

Partiendo de las definiciones anteriores, la probabilidad de que la secuencia y_1^T de observaciones ocurra conjuntamente con la secuencia de estados L en el modelo **m**, se obtiene como:

$$\begin{aligned}\mathbf{P(y,L/m)}&=\mathbf{b(y(1)/}i_1\mathbf{)b(y(2)/}i_2\mathbf{)...b(y(T)/}i_T\mathbf{)P(\underline{x}(1)=}i_1\mathbf{)a(}i_2/i_1\mathbf{)a(}i_3/i_2\mathbf{)....a(}i_T/i_{T-1}\mathbf{)}\\ &= \mathbf{P(\underline{x}(1)=}i_1\mathbf{)\ b(y(1)/}i_1\mathbf{)a(}i_2/i_1\mathbf{)b(y(2)/}i_2\mathbf{).....a(}i_T/i_{T-1}\mathbf{)b(y(T)/}i_T\mathbf{)}.\end{aligned} \tag{A.5}$$

Luego, la probabilidad de que la secuencia de observación y_1^T sea producida por el modelo **m** sería P(y/m) = $\sum_L$ P(y,L/m), lo que implica, la suma sobre todas las posibles secuencias de estados mutuamente excluyentes. Este cálculo es muy costoso desde el punto de vista del tiempo, lo que hace a este enfoque poco práctico, aun cuando se disponga de supercomputadores [1][17].

EL ENFOQUE BAUM-WELCH:

Este es un algoritmo que supera las dificultades encontradas en el enfoque anterior. Para mostrar su funcionamiento se definen las siguientes secuencias:

$y_{t_1}^{t_2}$ = {y(t_1),y(t_1+1), y(t_1+2), ..., y(t_2)}: Secuencia parcial de observaciones obtenida a partir de y_1^T que comienza en t = t_1 y termina en t = t_2.

y_1^t = {y(1), y(2), ..., y(t)}: Secuencia parcial hacia delante hasta el instante t.

y_{t+1}^T = {y(t+1), y(t+2), ..., y(T)}: Secuencia parcial hacia atrás desde el instante t+1.

La secuencia parcial hacia adelante hasta el instante T, es la secuencia de observaciones completa y_1^T.

Se definen también las probabilidades siguientes:

$\alpha(y_1^t,\mathbf{i})$ = P($y_{-1}^t = y_1^t$, x(t) = i /m) : Probabilidad conjunta de generar la secuencia parcial hacia adelante y_1^t y llegar al estado **i** en el t-ésimo paso o transición de estado, dado un **m**.

$\beta(y_{t+1}^T / i) = P(y_{-t+1}^T = y_{t+1}^T / \underline{x}(t) = i, m)$: Probabilidad de generar la secuencia parcial hacia atrás y_{t+1}^T, partiendo de la ocurrencia del estado **i** en el instante t, dado un **m**.

La figura A.3. permite entender el funcionamiento de los MOM, puesto que muestra las transiciones de estados que pueden ocurrir en cada instante de tiempo t.

La probabilidad de generar la secuencia y_1^{t+1}, para algún estado **j** en el instante t+1, si sólo hubiera una trayectoria hacia el estado j en t+1, que saliera del estado **i** en el paso t, se obtiene como sigue:

$$\alpha(y_1^{t+1}, j) = \alpha(y_1^t, i) P(\underline{x}(t+1) = j / \underline{x}(t) = i) P(\underline{y}(t+1) = y(t+1) / \underline{x}(t+1) = j)$$

$$= \alpha(y_1^t, i) a(j/i) b(y(t+1)/j) \qquad \text{(A.6)}$$

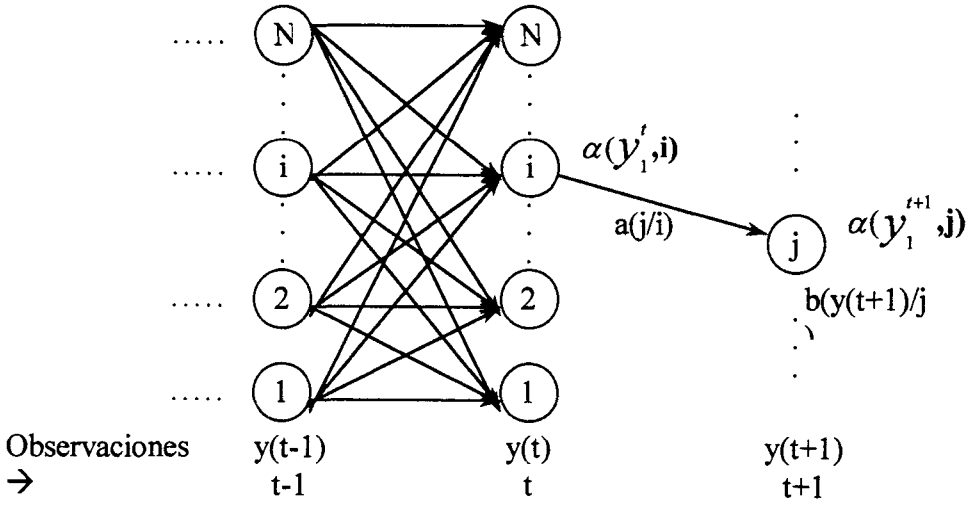


Figura A.3. Progreso del cambio de estados para generar las observaciones

Como en general, hay más de un estado i en el paso **t** a través del cual se puede llegar a **j,** en el paso t+1, entonces :

$$\alpha(y_1^{t+1}, j) = \sum_{i=1}^{N} \alpha(y_1^t, i) a(j/i) b(y(t+1)/j) \qquad \text{(A.7)}$$

La recursión se inicia con $\alpha(y_1^1, i) = P(\underline{x}(1) = i) b(y(1)/i)$ para cada **i.**

De la misma manera, trabajando con las secuencias hacia atrás (hacia la derecha en la figura A.3.) ,se obtienen las siguientes probabilidades:

$$\beta(y_{t+1}^T / i) = \sum_{j=1}^{N} \beta(y_{t+2}^T / j) a(j/i) b(y(t+1)/j) \qquad \text{(A.8)}$$

Para el cálculo de las probabilidades anteriores se trabaja en el rango $1 \leq t \leq T-1$.

Se usa la siguiente secuencia parcial inicial (ficticia) conveniente para arrancar la recursión [1][17][25]:

$\beta(y_{T+1}^{T}/i) = 1$, si **i** es un estado final legal (0, en otro caso).

Simplemente, es un recurso matemático para permitir el cálculo de la probabilidad de la ocurrencia de la secuencia hacia atrás.

Un estado final legal, es aquel o aquellos estados que según la topología que esté usando para los MOM, se seleccionen como estados en los cuales se puede generar la última observación de la secuencia y_1^T. Por ejemplo, en el modelo de izquierda a derecha de la figura A.2, el estado 5 sería claramente un estado final legal.

La probabilidad de que ocurra la secuencia y_1^T y que en un instante **t** pase por cualquier estado **i** se obtiene como:

$$P(y, \underline{x}(t) = i/m) = \alpha(y_1^t, i)\beta(y_{t+1}^T/i) \quad \text{(A.9)}$$

Finalmente, la verosimilitud P(y/m) (la medida de la probabilidad) se puede calcular en cualquier instante t de la manera siguiente:

$$P(y/m) = \sum_{i=1}^{N} \alpha(y_1^t, i)\beta(y_{t+1}^T/i) \quad \text{(A.10)}$$

En particular, si se trabaja en t=T, $P(y/m) = \sum_{Todoifinallegal} \alpha(y_1^T, i)$. Esta expresión es suficiente para encontrar la verosimilitud deseada.

Resumen del algoritmo F-B de Baum-Welch [1][17][25]:

Inicializar: $\alpha(y_1^1, j) = P(\underline{x}(1) = j)b(y(1)/j)$ **para j = 1,2,..,N**

Recursión : Para t = 1,...,T-1

Para j = 1,2,..,N

$$\alpha(y_1^{t+1}, j) = \sum_{i=1}^{N} \alpha(y_1^t, i)a(j/i)b(y(t+1)/j)$$

fin

fin

Final : **P(y/m)=** $\sum_{Todoifinallegal} \alpha(y_1^T, i)$

La técnica que se acaba de presentar, representa una reducción en varios órdenes de magnitud de la redundancia en los cálculos respecto al primer enfoque, lo que le da la propiedad para su uso en la realidad práctica, sin embargo, para poder llevarlo a la implementación se debe hacer una normalización de estas cantidades, para evitar pérdida de precisión numérica (lo que se conoce en el ambiente de los computistas como "underflow") al multiplicar una gran cantidad de valores que están entre cero y uno.

Más delante, se dedica una sección a explicar ese proceso de normalización.

A.9.2. SOLUCION AL PROBLEMA DE LA DESCODIFICACIÓN DE LOS MOM DE OBSERVACIONES DISCRETAS

Para resolver este problema hay que encontrar una secuencia de estados L = $\{i_1, i_2, ..., i_T\}$, tal que la probabilidad de ocurrencia de la secuencia de observaciones y = {y(1), y(2),...,y(t),...,y(T)}, a través de esa secuencia L, para un MOM **m** dado, sea mayor que la probabilidad obtenida desde cualquier otra secuencia de estados. Para obtener la mejor secuencia L existe un algoritmo famoso conocido como el **Algoritmo Viterbi** [1][17][25][26][57], que se explica a continuación.

EL ALGORITMO VITERBI:

Este algoritmo permite obtener P(y/m), buscando la mejor secuencia de estados posible a través de la cual, el MOM **m** pueda generar la secuencia de observaciones y_1^t dada, es decir, se calcula el número P(y, L^* /m).

Donde

$$L^* = \arg\max_L P(y, L^* / m) \qquad \text{(A.11)}$$

L^* : La mejor secuencia de estados.

y L es cualquier secuencia de estados de longitud T.

La técnica encuentra la mejor secuencia de estados concurrentemente con el cálculo de la verosimilitud P(y/m). Bajo esta óptica, el problema se reduce a un problema de optimización secuencial tratable con programación dinámica.

Obsérvese la figura A.4, en la cual se pinta en el eje de las abscisas los instantes t, en que ocurren las observaciones y en el eje de las ordenadas, los estados.

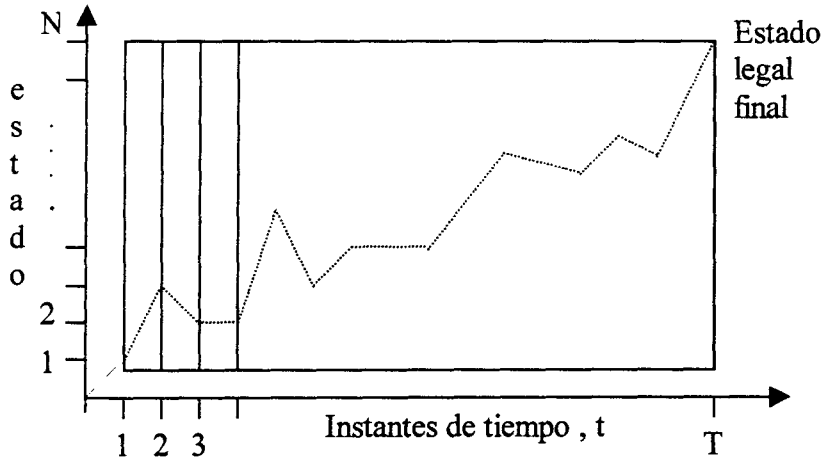


Figura A.4. Búsqueda de la mejor trayectoria de estados

La trayectoria punteada es la secuencia de estados que se busca. Dicha trayectoria avanza un tramo en cada paso, cada punto de dicha trayectoria es de la forma (t,i) y cada punto final legal es (T, i_f).

Para facilitar el cálculo de P(y, L^* /m) se definen las siguientes variables auxiliares:

- El costo tipo C ó costo de permanencia en cualquier punto (t,i):

$$C(t,i) = b(y(t)/i) = P(\underline{y} = y(t)/\underline{x}(t) = i) \qquad \text{(A.12)}$$

- El costo tipo Z ó costo de transición desde cualquier estado j al i:

$$Z((t,i)/(t-1,j)) = \text{a(i/j)} = \text{P}(\underline{x}(t) = i/\underline{x}(t-1) = j), \quad \text{con t} > 1 \text{ arbitrario.} \qquad \text{(A.13)}$$

Se supone que todas las trayectorias se originan en el punto ficticio (0,0) y hacen una transición a cualquier punto inicial (1,i). Al pasar al punto inicial, la trayectoria incurre en un costo Z de valor **a(i/0) = P($\underline{x}(1) = i$)** y en un costo C de valor **b(y(1)/i).**

- El costo tipo B ó costo acumulado al llegar a un punto cualquiera (t,i), desde el punto (t-1,j):

$$B[(t,i)/(t-1,j)] = Z[(t,i)/(t-1,j)]C(t,i)$$
$$= \text{a(i/j) b(y(t)/i)} \qquad \text{para t> 1.} \qquad \text{(A.14)}$$

$$B[(1,i)/(0,0)] = Z[(1,i)/(0,0)]C(1,i)$$
$$= \text{P}(\underline{x}(1)=\text{i}) \text{ b(y(1)/i)} \qquad \text{para t=1.} \qquad \text{(A.15)}$$

Considérese ahora una trayectoria completa de la forma : (0,0), $(1,i_1)$, $(2,i_2)$,...,(T,i_T).

Entonces, el costo total B asociado a dicha trayectoria será el producto de los costos B al pasar desde cada uno de sus puntos al siguiente:

$$B^{TA} = \prod_{t=1}^{T} B[(t, i_t)/(t-1, i_{t-1})]$$

$$= \prod_{t=1}^{T} a(i_t / i_{t-1}) b(y(t) / i_t) \qquad \text{(A.16)}$$

$a(i_1 / i_0) = a(i_1 / 0) = P(x(1) = i_1)$ se define así por conveniencia para arrancar el cálculo.

El costo total B de la trayectoria anterior es equivalente a la probabilidad de su ocurrencia conjunta con la secuencia de observación y_1^t.

Como la trayectoria que sea ha considerado por medio de puntos (t,i), es equivalente a la secuencia de estados L = $i_1, i_2, \ldots, i_T$, entonces se puede escribir que:

$$B^{TA} = \text{P(y,L/m)} \qquad \text{(A.17)}$$

Luego, la mejor trayectoria (mejor secuencia de estados), será la que maximice esa probabilidad, es decir, se tiene el interés de encontrar:

$$B^{TA^*} = P(y, L^* / m). \qquad \text{(A.18)}$$

Los productos involucrados para calcular esta probabilidad causan problemas computacionales debido a que pueden llegar a ser muy pequeños, sin embargo, esto se puede solucionar aplicando logaritmos negativos:

$$d[\ (t,i)/(t-1,j)\ ] = -\log B[\ (t,i)/(t-1,j)\ ] = -\log a(i/j) - \log b(y(t)/i). \qquad \text{(A.19)}$$

$$D = \sum_{t=1}^{T} d[\ (t,i)/(t-1,j)\ ] \qquad \text{(A.20)}$$

$$D = -\log B^{TA} \qquad \text{(A.21)}$$

Buscar la trayectoria con mínimo **D** es equivalente a buscar la trayectoria de máximo B^{TA}, puesto que, el logaritmo negativo de un número cercano a 1 es más pequeño que el logaritmo negativo de un número cercano a cero.

Si se quiere, el problema se puede interpretar como la búsqueda de la trayectoria más corta o lo que es lo mismo, la trayectoria que genera menor costo D, de la siguiente manera:

Si se define la distancia desde (0,0) hasta (t,i_t) sobre la mejor trayectoria como:

$$D_{min}(t,i_t)$$

Para t > 1, si se trabaja con los únicos predecesores a (t,i_t), de la forma $(t-1,i_{t-1})$ se tiene:

$$D_{min}(t,i_t) = \min_{(t-1,i_{t-1})} \{D_{min}(t-1,i_{t-1}) + d[(t,i_t)/(t-1,i_{t-1})]\}. \quad (A.22)$$

Como la minimización es sobre los estados previos, puesto que siempre se viene de t-1 :

$$D_{min}(t,i_t)= \min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) + d[(t,i_t)/(t-1,i_{t-1})]\} \quad \text{para } t > 1. \quad (A.23)$$

Al trabajar con los parámetros del modelo **m** se tiene:

$$D_{min}(t,i_t)= \min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) + [-\log a(i_t/i_{t-1}) - \log b(y(t)/i_t)]\} \quad \text{para } t > 1. \quad (A.24)$$

$D_{min}(0,0)=0$ y $a(i_1/0)$ es la probabilidad de ocurrencia del estado inicial i_1.

Al final de la búsqueda,

$$D^* = \min_{legal\, i_T} \{D_{min}(T,i_T)\} \quad (A.25)$$

será el logaritmo negativo de la probabilidad de ocurrencia conjunta de la secuencia de observación y_1^t, y la mejor secuencia de estados $L^*=i^*_1,....,i^*_T$ que la produce.

También se puede obtener el mejor de los estados finales como:

$$i^*_T = \arg\min_{i_T^{legal}} \{D_{min}(T,i_T)\}. \qquad (A.26)$$

$D^*=-\log P(y,L^*/m)$ es tan útil como la probabilidad misma para comparar Modelos Ocultos de Markov, por lo que no hay necesidad de hacer la conversión de nuevo al rango de las probabilidades [1][17]. Además, hay que tener en cuenta que los logaritmos pequeños son favorables porque implican probabilidades grandes.

Ahora, haciendo un procedimiento de backtracking como el que sigue, se puede recuperar la mejor secuencia de estados asociada con la D^* definiendo $\psi(t,i_t)$ como el penúltimo mejor estado de la trayectoria óptima parcial que termina en (t,i_t):

$$\psi(t,i_t) = \arg\min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) - \log a(i_t/i_{t-1}) - \log b(y(t)/i_t)\}$$

$$= \arg\min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) - \log a(i_t/i_{t-1})\} \qquad (A.27)$$

Los pasos del algoritmo que se ha revisado están basados en principios de programación dinámica. A veces, al algorítmo Viterbi para descodificar secuencias aleatorias, se le llama Forma estocástica de la Programación Dinámica debido a que los costos obtenidos son cantidades estocásticas [1].

Resumen del algorítmo Viterbi para calcular $P(y,L^*/m)$ ó una función de ésta [1][17][25][26][57]:

Inicialización : El origen de todas las trayectorias es el punto (0,0).

Para i=1,2,...,N

$D_{min}(1,i) = -\log a(i/0) - \log b(y(1)/i)$

$= - \log P(\underline{x}(1)=i) - \log b(y(1)/i)$

$\psi(1,i)=0$

Fin.

Recursión :

Para t=2,3,...,T

Para i_t=1,2,...,N

$$D_{min}(t,i_t) = \min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) - \log a(i_t/i_{t-1}) - \log b(y(t)/i_t)\} .$$

$$\psi(t,i_t) = \arg\min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) - \log a(i_t/i_{t-1}) - \log b(y(t)/i_t)\}$$

$$= \arg\min_{i_{t-1}} \{D_{min}(t-1,i_{t-1}) - \log a(i_t/i_{t-1})\}$$

Fin.

Fin.

Final : La distancia de la trayectoria óptima (0,0) a (T,i^*_T) está dada por :

$$D^* = \min_{i_T^{legal}} \{D_{min}(T,i_T)\}$$

La mejor secuencia de estados L^* se encuentra como sigue :

$$i^*_T = \arg\min_{i_T^{legal}} \{D_{min}(T,i_T)\}$$

Para t=T-1, T-2,...,1

$i^*_t = \psi(t,i^*_t)$

fin.

El algoritmo Viterbi hace menos cálculos que el algoritmo F-B de Baum-Welch para obtener D^*, lo que le da la propiedad de ser mucho más rápido que aquél.

A.9.3 SOLUCION AL PROBLEMA DEL ENTRENAMIENTO DE LOS MOM DE OBSERVACIONES DISCRETAS

El problema de entrenamiento o estimación de un MOM consiste en ajustar sus parámetros, partiendo de un conjunto de secuencias de observaciones, llamado **conjunto de entrenamiento**, tal que dicho conjunto sea representado por el modelo en la mejor forma posible, dependiendo del uso que se le desee dar en una aplicación particular. La idea detrás del modelo es que éste debe capturar características estadísticas comunes de esas secuencias.

Para ajustar o estimar los parámetros de un MOM se presentan los dos métodos siguientes:

EL METODO DE RE-ESTIMACION BAUM-WELCH PARA MOM DISCRETOS

Supóngase que se tienen secuencias de observaciones de la forma $y=y_1^T=\{y(1),...,y(T)\}$ y un MOM **m**.

El método supone un conjunto de parámetros iniciales para **m**, el cual se pretende mejorar para maximizar P(y/m), es decir, se parte de un MOM inicial cuyos parámetros se seleccionan aleatoriamente tomando en cuenta que [1][15][16][17][25][26]:

$$a(i/j) \geq 0, \quad 1 \leq i,j \leq N$$

y

$$\sum_{i=1}^{N} a(i/j) = 1, \quad 1 \leq j \leq N$$

$$b(k/i) \geq 0, \quad 1 \leq i \leq N, \quad 1 \leq k \leq K$$

y

$$\sum_{k=1}^{K} b(k/i) = 1,\ 1 \le i \le N$$

$$\prod(1) = [p(\underline{x}(1) = i] \ge 0,\ 1 \le i \le N$$

y

$$\sum_{i=1}^{N} p(\underline{x}(1) = i) = 1$$

El criterio de optimización por medio del cual el algoritmo de Baum-Welch maximiza P(y/m) se conoce como el **Criterio de Máxima Verosimilitud** y la función P(y/m) se conoce como una **Función de Verosimilitud** [1].

Para estimar los parámetros de un MOM se hacen las siguientes definiciones:

$u_{*/i}$: Conjunto de transiciones desde el estado i hacia todos los estados.

$u_{j/*}$: Conjunto de transiciones que entran al estado j desde todos los estados.

$\underset{-j}{y}(t)$: variable aleatoria que modela la observación emitida en el estado j en el instante t.

$\underline{u}$: proceso aleatorio con variables aleatorias $\underline{u}(t)$ que modelan las transiciones en cada instante t.

Luego, para una y_1^T y el modelo **m** se calculan las siguientes cantidades:

La probabilidad de estar en el estado i en el instante t y que ocurra una transición al estado j en t+1:

$\xi(i,j;t) = P(\underline{u}(t) = u_{j/i}/y,m)$, entonces recurriendo al teorema de Bayes, queda

$$= \frac{P(\underline{u}(t) = u_{j/i}, y/m)}{P(y/m)}$$

$$= \left\{ \frac{\alpha(y_1^t, i)a(j/i)b(y(t+1)/j)\beta(y_{t+2}^T/j)}{P(y/m)} \right\}, \text{ para } t=1,...,T-1. \qquad (A.28)$$

La probabilidad de que en el instante t ocurra una transición desde el estado i:

$$\gamma(i;t) = P(\underline{u}(t) \in u_{*/i}/y,m)$$

$$= \sum_{j=1}^{N} \xi(i,j;t)$$

$$= \left\{ \frac{\alpha(y_1^t, i)\beta(y_{t+1}^T/i)}{P(y/m)} \right\}, \text{ t=1,...,T-1} \qquad (A.29)$$

La probabilidad de que en el instante t, el MOM esté en el estado j:

$V(j;t) = P(\underline{x}(t) = j/y,m)$, de nuevo recurriendo al teorema de Bayes queda

$$= \frac{P(\underline{x}(t) = j, y/m)}{P(y/m)}$$

$$= \gamma(j;t), \qquad t=1,2,....,T-1$$

$$= \alpha(y_1^T, j), \qquad t=T$$

$$= \left\{ \frac{\alpha(y_1^t, j)\beta(y_{t+1}^T/j)}{P(y/m)} \right\}, \text{ t=1,...,T} \qquad (A.30)$$

La probabilidad de que estando en el estado j, en el instante t, se emita la observación ó símbolo k:

δ**(j,k ;t) = P(**$\underline{y}_j(t)$**=k /y,m)**, usando el teorema de Bayes

$$= \frac{P(y(t)=k, y/m)}{P(y/m)}$$

$= v(j;t)$, Si $y(t)=k$ y $t=1,...,T$

$$= \left\{ \frac{\alpha(y_1^t, j)\beta(y_{t+1}^T / j)}{P(y/m)} \right\}, \qquad \text{para } y(t)=k \text{ y } t=1,...,T \qquad (A.31)$$

Las cantidades antes señaladas, son cero para cualquier otro t que esté fuera de los rangos señalados.

Usando las definiciones anteriores se obtienen otras cantidades adicionales.

El número de transiciones desde el estado i al estado j durante la generación de la secuencia y(1),...,y(T).

$$\xi(i,j;*) = P(\underline{u}(*)=u_{j/i}/y,m) = \sum_{t=1}^{T-1} \xi(i,j;t) \qquad (A.32)$$

El número de transiciones desde el estado i durante la generación de la secuencia y(1),...,y(T).

$$\gamma(i;*) = P(\underline{u}(*) \in u_{*/i}/y,m) = \sum_{t=1}^{T-1} \gamma(j;t) \qquad (A.33)$$

El número de transiciones hacia el estado j durante la generación de la secuencia y(1),...,y(T).

$$v(j;*) = P(\underline{u}(*) \in u_{j/*}/y,m) = \sum_{t=1}^{T} v(j;t) \qquad (A.34)$$

El número de veces que la observación k y el estado j ocurren conjuntamente durante la generación de la secuencia y(1),...,y(T).

$$\delta(j,k;*) = P(\underset{-j}{y}(*) = k/y,m) = \sum_{t=1}^{T} \delta(j,k;t) = \sum_{\substack{t=1 \\ y(t)=k}}^{T} v(j;t) \qquad (A.35)$$

A partir de esas cantidades, los siguientes valores constituyen estimados razonables para los parámetros del MOM **m**:

$$\bar{a}(j/i) = \frac{\xi(i,j;*)}{\gamma(i;*)}$$

$$\bar{b}(k/j) = \frac{\delta(j,k;*)}{v(j;*)}$$

$$P(\underline{x}(1)=i) = \gamma(i;1)$$

$$\bar{a}(j/i) = \frac{\sum_{t=1}^{T-1} \alpha(y_1^t, i)a(j/i)b(y(t+1)/j)\beta(y_{t+2}^T/j)}{\sum_{t=1}^{T-1} \alpha(y_1^t, i)\beta(y_{t+1}^T/i)} \qquad (A.35)$$

$$\bar{b}(k/j) = \frac{\sum_{\substack{t=1 \\ y(t)=k}}^{T} \alpha(y_1^t, j)\beta(y_{t+1}^T/j)}{\sum_{t=1}^{T} \alpha(y_1^t, j)\beta(y_{t+1}^T/j)} \qquad (A.36)$$

$$P(\underline{x}(1)=i) = \frac{\alpha(y_1^1, i)\beta(y_2^T/i)}{P(y/m)} \qquad (A.37)$$

Al calcular estas tres cantidades para todo i, j y k, se tienen los parámetros de un nuevo modelo $\bar{m}$.

Se ha partido de un modelo **m** inicial, el cual se usa conjuntamente con la secuencia de entrenamiento y_1^T, para calcular cantidades con las cuales producir un nuevo modelo $\bar{m}$.

El objetivo ideal es usar la secuencia y_1^T de entrenamiento, para encontrar el mejor modelo m^* entre varios, es decir, se busca,

$$m^* = \arg\max_m P(y/m) \tag{A.38}$$

P(y/m) es una función no lineal de los parámetros del modelo **m** y contiene muchos máximos locales en el espacio multidimensional. En la figura A.5, se da una idea, en el espacio bidimensional, de cómo varía P(y/m).

La estimación repetida del modelo garantiza la convergencia a un **m** que corresponde a un máximo local de P(y/m), por lo tanto, por cada nuevo conjunto de parámetros se encuentra un $\bar{m}$ mejor que su antecesor, tal que $P(y/\bar{m}) > P(y/m)$. El procedimiento se detiene cuando se llega a un máximo local.

Esta re-estimación no produce el mejor modelo posible (el máximo global) por lo que se acostumbra correr el algoritmo varias veces con diferentes conjuntos de parámetros iniciales y tomar como el modelo entrenado, el **m** con el que se consiga el mayor valor de P(y/m).

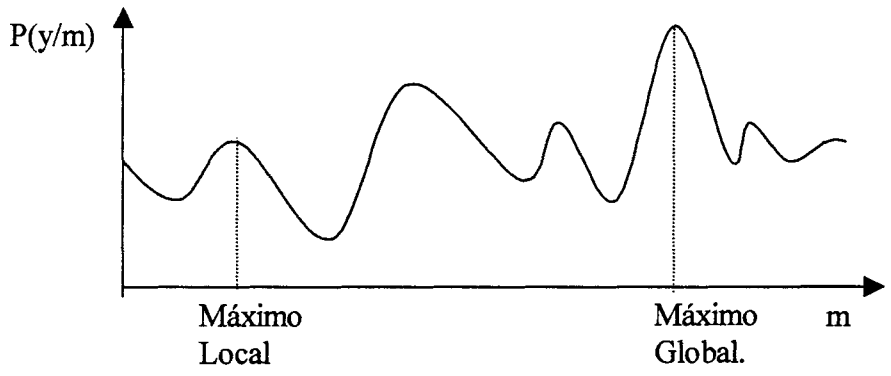


Figura A.5. Comportamiento de P(y/m) respecto a los parámetros de m

Resumen del algorítmo de entrenamiento Baum-Welch [1][17][25]:

Inicialización : Generar un modelo arbitrario **m**.

Recursión :

1.- Usar m={E,A,B, $\prod(1), \{y_k, 1 \le k \le K\}\}$ e y_1^T para calcular P(y/m) y las siguientes cantidades:

$$\xi(i,j;*) = P(\underline{u}(*)=u_{j/i}/y,m) = \sum_{t=1}^{T-1} \xi(i,j;t).$$

$$v(j;*) = P(\underline{u}(*) \in u_{j/*}/y,m) = \sum_{t=1}^{T} v(j;t)$$

$$\delta(j,k;*) = P(\underline{y}_j(*) = k/y,m) = \sum_{t=1}^{T} \delta(j,k;t) = \sum_{\substack{t=1 \\ y(t)=k}}^{T} v(j;t)$$

2.- Re-estimar el modelo (obtener los parámetros para el nuevo modelo $\bar{m}$) usando

$$\bar{a}(j/i) = \frac{\sum_{t=1}^{T-1} \alpha(y_1^t, i)a(j/i)b(y(t+1)/j)\beta(y_{t+2}^T/j)}{\sum_{t=1}^{T-1} \alpha(y_1^t, i)\beta(y_{t+1}^T/i)}$$

$$\bar{b}(k/j) = \frac{\sum_{\substack{t=1 \\ y(t)=k}}^{T} \alpha(y_1^t, j)\beta(y_{t+1}^T/j)}{\sum_{t=1}^{T} \alpha(y_1^t, j)\beta(y_{t+1}^T/j)}$$

$$P(\underline{x}(1)=i) = \frac{\alpha(y_1^1, i)\beta(y_2^T/i)}{P(y/m)}$$

3.- Se calcula la nueva $P(y/\bar{m})$.

4.- Si $P(y/\bar{m}) > P(y/m)$, es decir, si $P(y\backslash\bar{m}) - P(y\backslash m) > \varepsilon$ (ε es un valor suficientemente pequeño), se regresa al paso 1, con $m=\bar{m}$, de lo contrario se detiene el procedimiento.

Es importante recordar que este método funciona bien teóricamente, sin embargo, cuando se va a implementar en la práctica siempre se encuentran problemas de pérdida de precisión numérica, debido a que se trabaja con productos de probabilidades, por lo que para superar este problema hay que recurrir a una técnica razonable de escalamiento.

El algoritmo que se acaba de presentar trata una sola secuencia y_1^T, entonces para obtener un modelo que represente de forma más compacta a un proceso estocástico, es esencial entrenarlo con todo el conjunto de entrenamiento (con múltiples realizaciones del proceso). Para ello, hay que recordar que los numeradores y denominadores de los parámetros estimados de un modelo, representan el número de veces que ocurre algún evento a lo largo de la generación de una secuencia de observaciones. Luego, de acuerdo a esto simplemente se suman las ocurrencias de cada evento sobre todas las secuencias de observaciones del conjunto de entrenamiento. Bajo esa filosofía de trabajo, entonces:

$$\bar{a}(j/i) = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T-1}\alpha^s(y_1^t,i)a(j/i)b(y_{(t+1)}/j)\beta^s(y_{t+2}^T/j)}{\sum_{s=1}^{S}\sum_{t=1}^{T-1}\alpha^s(y_1^t,i)\beta^s(y_{t+1}^T/i)} \qquad \text{(A.39)}$$

$$\bar{b}(k/j) = \frac{\sum_{s=1}^{S}\sum_{\substack{t=1 \\ y(t)=k}}^{T}\alpha^s(y_1^t,j)\beta^s(y_{t+1}^T/j)}{\sum_{s=1}^{S}\sum_{t=1}^{T}\alpha^s(y_1^t,j)\beta^s(y_{t+1}^T/j)} \qquad \text{(A.40)}$$

$$P(\underline{x}(1)=i) = \frac{\sum_{s=1}^{S}\alpha^s(y_1^1,i)\beta^s(y_2^T/i)}{\sum_{s=1}^{S}P^s(y/m)} \qquad \text{(A.41)}$$

El super-índice s indica el resultado obtenido en el tratamiento de la s-ésima secuencia y_1^T, de las cuales hay S. Salvo este tratamiento, el procedimiento para el entrenamiento con múltiples secuencias es igual a cuando se trabaja con una sola realización del proceso estocástico [1][17][26].

EL METODO DE RE-ESTIMACION VITERBI PARA MOM DISCRETOS

Supóngase que se tiene un modelo **m** y una secuencia de entrenamiento y_1^T, a través de la cual se re-estiman los parámetros de un nuevo modelo. Se toma un estado cualquiera como estado inicial (normalmente el estado uno), de tal manera que las probabilidades de ocurrencia de estados en el instante t=1, no se requiere estimarlas.

El método comienza evaluando la P(y/m), usando la descodificación Viterbi que ya fue presentada [1][17][25]. En el recorrido que se sigue para obtener dicha probabilidad se almacenan las cantidades siguientes :

$n(u_{j/i})$: El número de transiciones desde el estado i al estado j.

$n(u_{*/i})$: El número de transiciones que parten del estado i.

$n(u_{j/*})$: El número de veces que se visita el estado j.

$n(\underset{-j}{y}(t)=k)$: El número de veces que la observación k y el estado j ocurren simultáneamente.

Luego, la re-estimación de los parámetros del nuevo modelo $\bar{m}$, se realiza con las ecuaciones siguientes:

$$\bar{a}(j/i) = \frac{n(u_{j/i})}{n(u_{*/i})} \tag{A.42}$$

$$\bar{b}(k/j) = \frac{n(\underset{-j}{y}(t)=k)}{n(u_{j/*})} \tag{A.43}$$

El procedimiento se detiene cuando se encuentra un $\bar{m}$ tal que $P(y/\bar{m}) < P(y/m)$, (en esta oportunidad hay que recordar que se trabaja con el concepto de costos ó con la trayectoria más corta).

Para la re-estimación a través de múltiples secuencias de observaciones y_1^T, se extienden las ecuaciones anteriores a:

$$\bar{a}(j/i) = \frac{\sum_{s=1}^{S} n^s(u_{j/i})}{\sum_{s=1}^{S} n^s(u_{*/i})} \tag{A.44}$$

$$\bar{b}(k/j) = \frac{\sum_{s=1}^{S} n^s(y_{-j}(t) = k)}{\sum_{s=1}^{S} n^s(u_{j/*})} \tag{A.45}$$

Igual que en Baum-Welch, el super-índice **s** indica el resultado obtenido en el tratamiento de la s-ésima secuencia y_1^T, de las cuales hay **S**.

Este enfoque es computacionalmente más eficiente que el método de Baum-Welch y presenta, hasta cierto punto, resultados comparables.

A.10. LOS MODELOS OCULTOS DE MARKOV DE OBSERVACIONES CONTINUAS

En el caso de reconocimiento de voz, los vectores de observaciones provienen de señales continuas, y aunque se pueden cuantificar esos vectores, como se hace cuando se trabaja con MOM de observaciones discretas, el proceso de cuantificación produce pérdida de información. Para evitar la pérdida de información es que se utilizan los MOM con funciones de densidades de probabilidad de observaciones continuas.

Para este tipo de Modelos Ocultos de Markov la estructura matemática es de la forma [1][17][25]:

$$m = \{\mathbf{E}, \prod(1), \mathbf{A}, \{ f_{\underline{y}/\underline{x}}(\xi/i), 1 \le i \le N \}\} \tag{A.46}$$

donde $f_{\underline{y}/\underline{x}}(\xi/i)$ es la función de densidad multivariante que caracteriza la generación de las observaciones en el estado **i**.

A.10.1. SOLUCION A LOS PROBLEMAS DE EVALUACIÓN Y DE DESCODIFICACIÓN DE LOS MOM CONTINUOS

La diferencia esencial, con respecto a cualquiera de los enfoques presentados para el caso de MOM de observaciones discretas estriba en que ahora, se define [1][17][25]

b(y(t)/i) = $f_{\underline{y}/\underline{x}}(y(t)/i)$ para cualquier observación y(t)

por lo tanto, se pueden implementar esos algoritmos, con esa nueva b(y(t)/i).

A.10.2. SOLUCION AL PROBLEMA DE ENTRENAMIENTO DE LOS MOM CONTINUOS

La función de densidad de probabilidad que se utiliza con mayor frecuencia, en reconocimiento es una mezcla de gausianas, para generar las observaciones en los estados de los MOM de observaciones contínuas [1][17][25]. Esa función tiene la forma siguiente:

$$f_{\underline{y}/\underline{x}}(\xi/i) = \sum_{m=1}^{M} c_{im} n(\xi; \mu_{im}, C_{im})$$

$$= \sum_{m=1}^{M} c_{im} \left(\frac{1}{\sqrt{(2\Pi)^{P} Det\, C_{im}}} e^{-\frac{1}{2}(\xi-\mu_{im})^{T} C_{im}(\xi-\mu_{im})} \right) \qquad \text{(A.47)}$$

donde c_{im} es el coeficiente o peso de la m-ésima componente gausiana del estado i, $n(\xi; \mu_{im}, C_{im})$ es una función de densidad de probabilidad gausiana multivariante de media μ_{im} y matriz de covarianza C_{im}, ξ es el vector de observaciones que se está modelando y P es el tamaño de ese vector.

La mezcla de gausianas $f_{\underline{y}/\underline{x}}(\xi / i)$, es una función de densidad de probabilidad normalizada en el sentido de que los coeficientes de sus componentes son no negativos y cumplen la restricción siguiente:

$$\sum_{m=1}^{M} c_{im} = 1 \qquad \text{para i = 1, 2, ..., N} \tag{A.48}$$

Cuando M es un número suficientemente grande, entonces se puede utilizar a $f_{\underline{y}/\underline{x}}(\xi / i)$ para modelar cualquier función de densidad de probabilidad [1], (la m que aparece aquí se refiere a una función componente entre M gausianas, mientras que la m que se ha venido utilizando, en las secciones precedentes, se refiere a los MOM, a los que se han llamado modelos m).

EL METODO DE RE-ESTIMACION BAUM-WELCH PARA MOM CONTINUOS :

Para este tipo de re-estimación se definen las cantidades siguientes: c_{iL}, μ_{iL} y C_{iL} para la función de densidad de probabilidad **L** (componente **L**) del estado **i**. También se define la probabilidad de que en el instante **t**, ocurra el estado **i** cuando la función de densidad de probabilidad **L** genera la observación **y(t)** [1][17] (en este caso una observación es un vector).

$v(i;t,L) = P(\underline{x}(t) = i / y(t))$

$$= \frac{\alpha(y_1^t, i)\beta(y_{t+1}^T / i)}{\sum_{t=1}^{N} \alpha(y_1^t, j)\beta(y_{t+1}^T / j)} \times \frac{c_{iL}\, n(\xi; \mu_{iL}, C_{iL})}{\sum_{m=1}^{M} c_{im}\, n(\xi; \mu_{im}, C_{im})} \tag{A.49}$$

Ahora, se define

$$v(j;*,L) = \sum_{t=1}^{T} v(j;t,L) \tag{A.50}$$

Con estas cantidades se obtienen, para cada estado **i** y por cada componente **L**, los valores estimados de los coeficientes de las funciones de densidad de probabilidad, sus vectores de medias y sus matrices de covarianzas:

$$\bar{c}_{iL} = \frac{v(i;*,L)}{\sum_{m=1}^{M} v(i;*,m)} \tag{A.51}$$

$$\bar{\mu}_{iL} = \frac{\sum_{t=1}^{T} v(i,t,L)y(t)}{v(i;*,L)} \tag{A.52}$$

$$\bar{C}_{iL} = \frac{\sum_{t=1}^{T} v(i;t,L)(y(t)-\mu_{iL})(y(t)-\mu_{iL})^T}{v(i;*,L)} \tag{A.53}$$

La ecuación (A.51) representa la razón entre, el número de veces que la secuencia de estados pasa por el estado i y la L-ésima gausiana genera la secuencia de observaciones y_1^T; y el número de veces que la secuencia de estados pasa por el estado **i**, durante la generación de la secuencia de observaciones y_1^T, por parte de la función $f_{y/x}(y_1^T/i)$.

En este caso, y_1^T es una secuencia de vectores sin cuantificar, mientras que en el caso de MOM de observaciones discretas, se trata de una secuencia de valores (no vectores) que son el resultado de un proceso de cuantificación.

La ecuación (A.52) representa el vector promedio de los vectores de observaciones, ponderado de acuerdo a la verosimilitud producida por la componente **L** en el estado i.

El cálculo de la covarianza estimada, es también la matriz de covarianza de los vectores de observaciones, ponderada de acuerdo a la verosimilitud producida por la componente **L** en el estado i.

En reconocimiento, los vectores iniciales de medias y de covarianzas de las componentes L para cada estado i, se obtienen de todos los vectores de las señales de entrenamiento.

A excepción de estos cálculos, el algoritmo es el mismo que en el caso de los MOM de observaciones discretas.

EL METODO DE RE-ESTIMACION VITERBI EN EL CASO DE MOM CONTINUOS

En el enfoque Viterbi, los vectores de medias y las matrices de covarianzas para las densidades de las observaciones se re-estiman por promedios y luego, se trabaja como en el caso discreto pero con b(y(t)/i) = $f_{\underline{y}/\underline{x}}(y(t)/i)$.

Supóngase que en un estado **i** hay una sola gausiana para generar las observaciones. En este caso, cada vector de observaciones se asigna al estado que lo produjo en la trayectoria óptima cuando se examina la información por backtraking.

Si el vector de observaciones y(t) es producido por el estado **i**, entonces "y(t)→ i" [1], y supóngase que N_i vectores de observaciones se asignan al estado **i**, entonces

$$\mu_i = \frac{1}{N_i} \sum_{\substack{t=1 \\ y(t)-->i}}^{T} y(t) \qquad \text{(A.54)}$$

$$C_i = \frac{1}{N_i} \sum_{\substack{t=1 \\ y(t)-->i}}^{T} (y(t)-\mu_i)(y(t)-\mu_i)^T \qquad \text{(A.55)}$$

Cuando hay M > 1 gausianas por estado, el conjunto de los vectores de observaciones asignados a cada estado se subdivide en M subconjuntos, y luego se estiman los vectores de medias y las matrices de covarianzas de cada subconjunto respectivamente. Este proceso se puede efectuar por medio del algoritmo de las K-medias [1], con K=M.

Si hay N_{iL} vectores asignados a la L-ésima componente gausiana en el estado **i**, entonces el coeficientes de esa componente c_{iL}, se re-estima como

$$c_{iL} = \frac{N_{iL}}{N_i} \tag{A.56}$$

A.11. ESCALAMIENTO DE LAS PROBABILIDADES EN EL ALGORITMO BAUM-WELCH

Se realiza una normalización, para $\alpha(y_1^t,\text{i})$ y $\beta(y_{T+1}^T / i)$.

En el instante t = 1 se calcula el factor de normalización [1][17]

$$c_1 = \frac{1}{\sum_{j=1}^{N} \alpha(y_1^1, j)} \tag{A.57}$$

Luego se escala $\alpha(y_1^1,\text{i})$ como

$$\bar{\alpha}(y_1^1,\text{i}) = c_1\, \alpha(y_1^1,\text{i}) \qquad \text{para i} = 1,\ldots,N \tag{A.58}$$

Luego en t = 2,

$$\begin{aligned}
\tilde{\alpha}(y_1^2, i) &= \sum_{j=1}^{N} \bar{\alpha}(y_1^1,\text{j})\ \text{a(i/j)}\ \text{b(y(2)/i)} \qquad \text{para i} = 1, \ldots, N \\
&= \sum_{j=1}^{N} c_1\, \alpha(y_1^1,\text{j})\ \text{a(i/j)}\ \text{b(y(2)/i)} \\
&= c_1 \sum_{j=1}^{N} \alpha(y_1^1,\text{j})\ \text{a(i/j)}\ \text{b(y(2)/i)} \\
&= c_1\, \alpha(y_1^2,\text{i})
\end{aligned} \tag{A.59}$$

por lo tanto, la normalización en t = 2 se hará como

$$c_2 = \frac{1}{\sum_{j=1}^{N} \tilde{\alpha}(y_1^2, j)} \tag{A.60}$$

$$\bar{\alpha}(y_1^2,\mathbf{i}) = c_2\,\tilde{\alpha}(y_1^2,i)$$

$$= c_2 c_1\ \alpha(y_1^2,\mathrm{i}) \qquad \text{(A.61)}$$

al proceder de esta manera recursiva, para t = 2, ...,T se tiene la siguiente forma general de escalar esas cantidades:

$$\bar{\alpha}(y_1^t,\mathbf{i}) = c_t\,\tilde{\alpha}(y_1^t,i) \qquad \text{para i= 1, ..., N} \qquad \text{(A.65)}$$

Para calcular la P(y/m) escalada, se recurre a la propiedad

$$\bar{\alpha}(y_1^T,\mathbf{i}) = \prod_{t=1}^{T} c_t\,\alpha(y_1^T,i) \qquad \text{(A.66)}$$

por lo tanto

$$\sum_{i=1}^{N}\ \bar{\alpha}(y_1^T,\mathrm{i}) = 1 \qquad \text{(A.67)}$$

$$\sum_{i=1}^{N}\ \prod_{t=1}^{T} c_t\,\alpha(y_1^T,i) = 1 \qquad \text{(A.68)}$$

$$\prod_{t=1}^{T} c_t \sum_{i=1}^{N}\ \alpha(y_1^T,i) = 1$$

Sabemos que $P(y/m) = \sum_{i=1}^{N}\ \alpha(y_1^T,i)$ entonces

$\prod_{t=1}^{T} c_t P(y/m) = 1$, de aquí se obtiene

$$P(y/m) = \frac{1}{\prod_{t=1}^{T} c_t} \tag{A.69}$$

O en términos de logaritmos, como

$$\bar{P}(y/m) = \mathbf{Log(P(y/m))} = -\sum_{t=1}^{T} \log c_t \tag{A.70}$$

En cuanto al escalamiento de las cantidades $\beta(y_{t+1}^{T} / i)$ el proceso es más sencillo, puesto que, se utilizan los mismos coeficientes que se emplean en la normalización de las cantidades $\alpha(y_1^t, i)$. Aunque pueda parecer que escoger los mismos valores de normalización, no garantice que el rango de los valores para los $\beta(y_{t+1}^{T} / i)$ escalados sean adecuados, lo cierto es que esos valores quedan acotados, que es lo que realmente se busca, para evitar la pérdida de precisión en los cálculos.

Para el instante t = T,

$$\bar{\beta}(y_t^T / i) = c_T \, \beta(y_t^T / i) \tag{A.71}$$

mientras que para t = T-2, ...,1

$$\bar{\beta}(y_{t+1}^T / i) = c_t \, \tilde{\beta}(y_{t+1}^T / i) \tag{A.72}$$

$$\text{con } \tilde{\beta}(y_{t+1}^T / i) = \sum_{j=1}^{N} \bar{\beta}(y_{t+2}^T / j) a(j/i) b(y(t+2)/j) \tag{A.73}$$

ANEXO B

PARAMETRIZACION DE LAS SEÑALES DE VOZ A TRAVES DE ANÁLISIS DE PREDICCIÓN LINEAL Y CEPSTRAL

B.1. INTRODUCCION

La señal de voz generada por el aparato articulatorio humano es una onda de presión que para que pueda ser procesada hay que convertirla a una señal digital. Primero se transforma la onda de presión en una señal eléctrica analógica, y luego de analógica a digital, sobre la que finalmente, se aplican técnicas de procesamiento digital de señales.

En reconocimiento de voz es muy importante encontrar una representación paramétrica de la señal que contenga además de información instantánea de la señal, información sobre la variación temporal de la propiedades de la señal. El objetivo de esta representación paramétrica es la extracción de las características acústicas más relevantes para el proceso de reconocimiento.

En la literatura [7][8][11][57][58], se encuentra que un tipo de parámetros óptimo para reconocimiento de voz son los coeficientes cepstrales. Los coeficientes cepstrales son un conjunto de parámetros apropiado para representar las características espectrales de la señal de voz; sin embargo, sólo contienen información instantánea de la señal. La incorporación de características espectrales dinámicas fue sugerida por Furui, [7] incorporando los coeficientes que representan las primeras derivadas de los coeficientes cepstrales. Otros trabajos [48][58] han mostrado que al aumentar el vector de parámetros añadiendo la energía así como su derivada se contribuye al reconocimiento.

El primer paso de la parametrización es dividir la señal de voz de entrada en segmentos o tramas, y a partir de cada segmento hay que realizar una estimación espectral suavizada. Los segmentos se van solapando, de modo que la separación entre ellos sea de unos 10

milisegundos, y la longitud de los segmentos no sea superior a 30 milisegundos. La señal de cada segmento se multiplica por una ventana Hamming para atenuar los efectos de límite de la ventana cuadrada, y se realiza un preénfasis con la finalidad de compensar la atenuación a altas frecuencias producida por la radiación de los labios [17][26][57][58].

La estimación espectral requerida se puede realizar, mediante análisis de Predicción Lineal, o mediante análisis de Fourier. En cualquiera de los casos hay que hacer varias transformaciones adicionales a las mencionadas para obtener los vectores de parámetros acústicos.

Es importante tener presente que el análisis de la voz se hace en tiempo finito, es decir, sobre tramos de la señal y que la voz es un proceso dinámico altamente cambiante por lo que la longitud de los tramos se selecciona sobre algunas decenas de milisegundos para garantizar que la señal presente algún grado de estacionaridad.

El objetivo del análisis LP y Cepstral es determinar, cómo esos tramos se comportan en el tiempo, por lo que se trata de capturar propiedades transitorias de la señal. Existen muchas razones por las cuales en procesamiento de voz es importante realizar un análisis por segmentos de la señal, debido a que de esta manera se pueden identificar pronunciaciones cómo de vocales, de consonantes, de sonidos "sonoros", sonidos "sordos", de fonemas, la energía temporal y espectral, etc., [1][8][39].

Las técnicas de parametrización LP y Cepstral han sido la base de muchos resultados teóricos y prácticos, por lo que según Deller et al [1], es difícil concebir la tecnología del habla moderna sin éstas.

B.2. ANALISIS DE PREDICCION LINEAL (LP)

El análisis de predicción lineal se basa en la suposición de un modelo de producción de voz que se puede encontrar descrito en la literatura específica relacionada con el tema, como por ejemplo en [1], que proporciona un sistema de análisis para las señales de voz, donde se plantea que a través de muchos estudios en el campo de la tecnología del habla, se han originado diversos modelos para el proceso de producción de la voz, hasta llegar a considerar

que durante un segmento estacionario o cuasi-estacionario de la señal s(n), un "buen" modelo estaría dado por una función de transferencia con polos y ceros, de la siguiente forma:

$$\Theta(z) = \Theta_0^1 \frac{1+\sum_{i=1}^{L} b(i) Z^{-i}}{1-\sum_{i=1}^{R} a(i) Z^{-i}} \qquad \text{(B.1)}$$

Este modelo estaría activado por una secuencia de excitación e(n), que tiene la forma siguiente:

e(n) = $\sum_{q=-\infty}^{\infty} \delta$ (n-qp), para la producción de sonidos del tipo "sonoros" y sería un ruido no correlacionado de media cero y varianza uno, para la producción de sonidos "sordos".

δ (n), es la respuesta a un tren de impulsos de período **p**, generado por la glotis.

En la figura B.1, se muestra un diagrama de bloques para el sistema de producción de la voz descrito a través de esta función de transferencia.

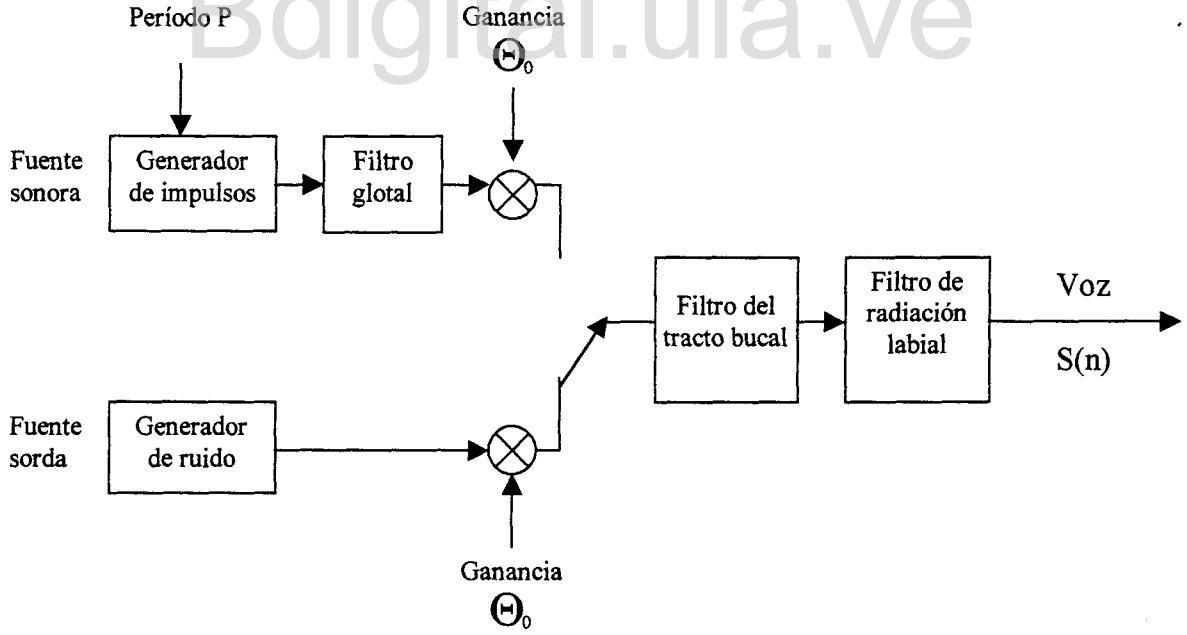


Figura B.1. Modelo de la producción de voz en tiempo discreto.

El objetivo del análisis LP es identificar los parámetros de una función de transferencia todo polo que constituye un estimado del sistema de producción de voz dado por (B.1):

$$\bar{\Theta}(z) = \frac{1}{1 - \sum_{i=1}^{M} \bar{a}(i) Z^{-i}} \qquad \text{(El modelo LP)} \qquad \text{(B.2)}$$

donde $\bar{\Theta}(z)$ constituye el modelo estimado, mientras que $\Theta(z)$, se supone que es el sistema real.

En la figura B.2, se muestra un diagrama de bloques para el sistema de producción de la voz estimado, descrito a través de la función de transferencia (B.2).

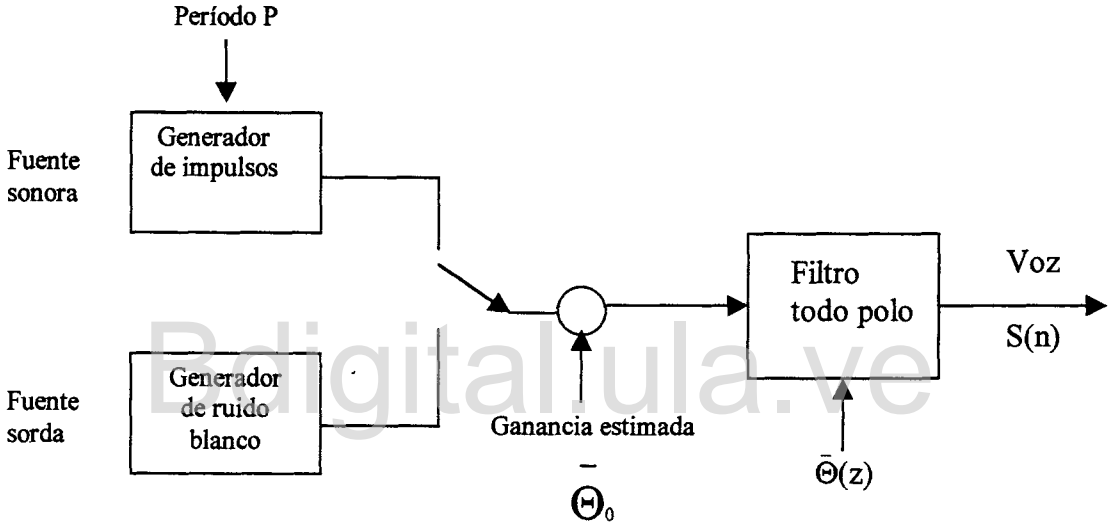


Figura B.2. Modelo de producción de voz utilizado por el análisis LP.

La justificación de por qué se usa un modelo todo polo, si se sabe que el modelo más adecuado debería ser de polos y ceros, se debe más que todo a la dificultad de tratar tal problema en forma analítica. Mientras que en el caso del sistema todo polos, la búsqueda de los parámetros se realiza, a través de ecuaciones lineales, trabajando sólo con valores pasados de la salida y donde se requiere mínima información de la entrada (sólo el valor presente).

Otra justificación, si se quiere más realista, es el hecho de que la información más significativa de la señal acústica se puede extraer de la magnitud espectral, y que la magnitud espectral se puede modelar en forma exacta con polos estables, no así la fase, pero las relaciones de fase entre las componentes de la voz no tienen efecto sobre la percepción de ésta, vale decir, por

ejemplo, podemos entender claramente la voz de un locutor que esté parado en algún sitio de una casa, si nos movemos de cuarto en cuarto. Con esto queremos decir que aunque las relaciones de fase entre las componentes de la voz estén cambiando, la voz parece la misma si se le da suficiente amplitud. En este sentido, el modelo LP puede preservar de forma exacta la dinámica de la magnitud espectral (la información) de la voz, pero no puede retener las características de fase [1]. De hecho, la naturaleza todo polo, estable, de la representación LP condiciona al modelo a ser de fase mínima. Si el objetivo es codificar, almacenar, sintetizar la voz a partir de las características de la magnitud espectral (sin tomar en cuenta la dinámica temporal), entonces el modelo LP planteado es perfectamente válido y útil.

A partir de los coeficientes de predicción lineal se pueden obtener los coeficientes cepstrales como se muestra en [1] y [17].

B.2.1. JUSTIFICACIÓN MATEMÁTICA PARA EL MODELO LP

La justificación del modelo de predicción lineal se encuentra en los dos lemas siguientes [1]:

Lema 1: Cualquier sistema causal racional de la forma (B.1) puede descomponerse como:

$$\Theta(z) = \Theta_0 \; \Theta_{min}(z) \; \Theta_{ap}(z) \qquad \text{(B.3)}$$

donde $\Theta_{min}(z)$ es un componente de fase mínima y $\Theta_{ap}(z)$, es todo paso, es decir, $\left|\Theta_{ap}(e^{jw})\right| = 1 \quad \forall$ w (en el plano de la frecuencia)

y Θ_0 es una constante relacionada con Θ_0^1 y con las singularidades de Θ(z).

Lema 2: El componente de fase mínima que resulta de la descomposición indicada por el lema 1, se expresa como el sistema todo polo siguiente:

$$\Theta_{min}(z)= \frac{1}{1-\sum_{i=1}^{I} a(i) Z^{-i}} \qquad (B.4)$$

donde I, en la práctica es un entero relativamente pequeño, aun cuando teóricamente puede ser infinito.

De acuerdo a los lemas 1 y 2, el modelo de Predicción Lineal (LP) representa la porción todo polo y de fase mínima de Θ(z), por lo tanto, el modelo real se puede expresar como:

$$\Theta(z)= \Theta_0 \frac{1}{1-\sum_{i=1}^{I} a(i) Z^{-i}} \Theta_{ap}(z) \qquad (B.5)$$

y $\Theta_{min}(z)= \frac{1}{1-\sum_{i=1}^{I} a(i) Z^{-i}}$, será el modelo al que se le identifican sus parámetros.

Al hacer I = M como en la expresión (B.2), entonces la tarea consiste en buscar los $\bar{a}(i)$ de tal forma que sean buenos estimados de los a(i) del modelo (B.5), con lo cual se producirá un buen modelo en cuanto a magnitud espectral para Θ(z), ya que $|\Theta(w)| = \Theta_0 |\Theta_{min}(w)|$.1 ∀ w.

B.2.2. ECUACIONES DE PREDICCIÓN LÍNEAL

Debido a que los parámetros $\bar{a}(i)$ estimados constituyen una representación paramétrica o un código para la señal acústica que contiene información de la magnitud espectral de dicha señal, resulta más eficiente transmitir, almacenar, analizar, sintetizar o reconocer esa cantidad M de parámetros, que trabajar con la gran cantidad de datos de donde se derivan esos parámetros.

De la ecuación (B.5), la señal de voz será:

$$S(z) = \Theta(z)E(z) = \Theta_0 \Theta_{min}(z) E^1(z) \qquad (B.6)$$

Donde S(z) es la transformada Z de la secuencia de voz s(n) y E(z) es la transformada Z de la secuencia de excitación e(n), con

$$E^1(z) = E(z)\,\Theta_{ap}(z), \text{ donde } \Theta_{ap}(z)=1 \qquad (B.7)$$

Por lo tanto en el dominio del tiempo:

$$S(n) = \sum_{i=1}^{M} a(i)S(n-i) + \Theta_0 e^1(n)$$

$$= \mathbf{a}^T\mathbf{S(n)} + \Theta_0 \mathbf{e}^1\mathbf{(n)} \qquad (B.8)$$

La ecuación (B.8) es la representación vectorial de la secuencia de voz s(n), es decir,

$$\mathbf{a}=[\mathbf{a(1)\ a(2)\ ....\ a(M)}]^T$$
$$\mathbf{S(n)} = [\mathbf{S(n-1)\ S(n-2)\ ....\ S(n-M)}]^T$$

Donde S(n) es la salida del componente de fase mínima de Θ(z) que es excitado por una versión de e(n). Excepto por este término de entrada, S(n) puede predecirse usando una combinación lineal de sus M valores pasados (en términos estadísticos se dice que la salida regresa sobre si misma y el modelo se llama el **modelo autoregresivo AR(M)**).

Entonces estamos frente a un modelo **AR(M)** con M parámetros o polos, o simplemente, frente a un Modelo AR de orden M.

Hay varias formas de calcular los $\bar{a}(i)$ dependiendo de las distintas interpretaciones que se le den al análisis LP, es decir, por ejemplo, se puede interpretar el análisis LP como un problema de identificación de sistemas, como un problema de filtraje inverso, como un problema de análisis espectral o como un problema de predicción lineal [1].

B.2.2.1. Cálculo de los parámetros a través del enfoque de Predicción Lineal

El problema se reduce a la búsqueda de un filtro predictivo FIR (filtro de respuesta impulsiva finita), que minimiza el cuadrado promedio del error de predicción como lo muestra la figura B.3.

Si $\bar{S}(n) = \sum_{i=1}^{M} \bar{a}(i)\, S(\eta\text{-}i)$ se ve como la predicción de S(n) entonces, $P(z) = \sum_{i=1}^{M} \bar{a}(i) Z^{-i}$ es el filtro de predicción, donde $\bar{e}(n)$ es el error de predicción, (se observará que se trata de un trabajo con filtros inversos).

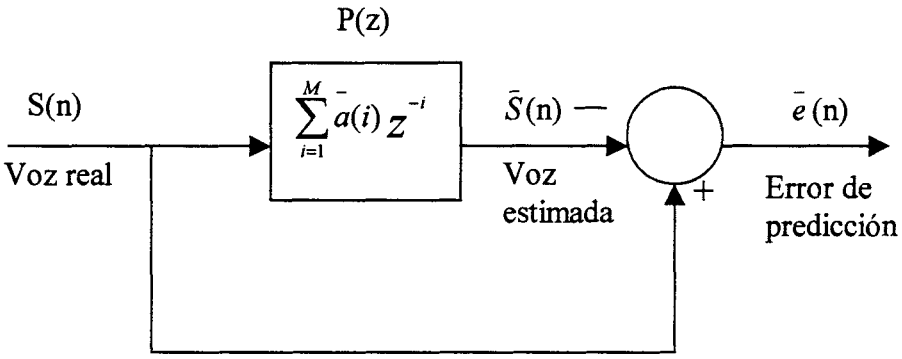


Figura B.3. Filtraje inverso para obtener los parámetros LP.

B.2.3. Regla aceptada para fijar el orden del modelo LP

En la práctica, el orden real, M, del sistema es por supuesto desconocido. Sin embargo, en la actualidad se acepta como regla para escoger el orden del modelo, la siguiente [1][17]:

$$M = \begin{cases} Fs + (4 \text{ o } 5)\text{ , en el caso sonoro} \\ Fs\text{ , en el caso sordo.} \end{cases}$$

Donde Fs es la frecuencia de muestreo de la señal acústica en Khz.

B.2.4. El análisis LP por tramos de la señal acústica

Existen dos métodos muy conocidos para obtener la solución del problema LP: el método de la Autocorrelación y el método de la Covarianza [1][39].

En este anexo se presenta sólo el método de la autocorrelación, porque fue dicho método a través del cual se programó un módulo de parametrización que podría formar parte de un futuro reconocedor venezolano.

B.2.4.1. El Método de la autocorrelación

Se desea estimar los parámetros LP sobre un segmento de N puntos que terminan en el instante m, es decir, se trabaja sobre el rango S(m-(N-1)), S(m-(N-2)),...,S(m-(N-N)) $\Rightarrow$ S(m-N+1), S(m-N+2),...,S(m).

Los estimadores naturales para los parámetros LP que usan sólo los datos del segmento especificado, son los estimadores, $r_s(\eta;m)$, de la autocorrelación [1].

A continuación se presentan las ecuaciones correspondientes a los M parámetros de autocorrelación del tramo que se está analizando:

$$\sum_{i=1}^{M} \bar{a}(i) r_s(\eta - i) = r_s(\eta;m), \qquad \eta = 1,2,....,M \tag{B.9}$$

y

$$\mathbf{R_s(m)}\ \bar{a}\mathbf{(m)} = \mathbf{r_s(m)} \quad \Rightarrow \quad \bar{a}\mathbf{(m)} = R_s^{-1}\mathbf{(m)}\ \mathbf{r_s(m)}, \text{ es su forma matricial.} \tag{B.10}$$

Donde $R_s(m)$ es la matriz de autocorrelación que constituye un operador Toeplitz, puesto que los elementos de cualquiera de sus diagonales son iguales.

$\bar{a}\mathbf{(m)}$ **y** $\mathbf{r_s(m)}$ son vectores de los parámetros estimados LP y de autocorrelación respectivamente, asociados con la secuencia S(n) en el tramo.

La solución a estas ecuaciones se conoce como el método de la autocorrelación para determinar los coeficientes $\bar{a}(i)$.

B.2.4.2. Método para resolver las ecuaciones de Autocorrelación

El método de la autocorrelación está relacionado con la solución de un problema lineal de la forma **Ax = b**, donde **A** es una matriz y donde **x** y **b** son vectores (sistemas de ecuaciones lineales para los cuales existen muchas técnicas de solución), además, el método de la autocorrelación produce una **A** simétrica y Toeplitz.

En procesamiento de señales de voz se recurre con mucha frecuencia a la técnica conocida como el algoritmo Levinson-Durbin para resolver éste problema, cuando se usan los parámetros de la autocorrelación para obtener los parámetros LP [1][17]. Por esta razón, a continuación se describe esa técnica:

En 1947 Levinson [1], publicó un algoritmo para resolver el problema **Ax=b**, en el cual **A** es Toeplitz, simétrica, y definida positiva, y b es arbitrario (las ecuaciones de autocorrelación son exactamente de esta forma). En 1960, Durbin publicó un algoritmo ligeramente más eficiente para este caso, dando origen al algoritmo conocido como algoritmo Levinson-Durbin (L-D).

La solución de las ecuaciones de autocorrelación para el modelo predictor de orden M, la obtiene la recursión L-D en forma sucesiva, es decir, se trabaja partiendo de obtener los parámetros de modelos de orden menor (de orden cero hacia arriba) hasta alcanzar los parámetros para un modelo de orden M.

La forma matricial de las ecuaciones de autocorrelación dada por (B.10) se puede re-escribir como:

$$- R_s(m)\ \bar{a}^M(m) + r_s(m) = 0 \qquad \text{(B.11)}$$

Si E(n;m) es la secuencia de predicción residual generada debido al estimado de los $\bar{a}(m)$ entonces

$$r_s(0;m) - \sum_{i=1}^{M} \bar{a}^M(i;m)\, r_s(i;m) = \frac{1}{N}\sum_{n=-\infty}^{\infty} E(n;m)^2 \qquad \text{(B.12)}$$

El término a la derecha de (B.12) se puede interpretar como la energía de la secuencia de predicción residual escalada por (1/N), producida por el filtro inverso de orden M, y se simboliza como $\varepsilon^M(m)$.

Si se construye una nueva estructura matricial utilizando (B.11) y (B.12).

$$\begin{bmatrix} r_s(0;m) & \mathbf{r}_s^T(m) \\ \mathbf{r}_s(m) & \mathbf{R}_s(m) \end{bmatrix} \begin{bmatrix} 1 \\ -\bar{a}(m) \end{bmatrix} = \begin{bmatrix} \varepsilon^M(m) \\ \mathbf{0} \end{bmatrix} \qquad \text{(B.13)}$$

donde $\mathbf{R}_s(m)$ es una matriz y $\mathbf{r}_s^T(m)$, $\mathbf{r}_s(m)$, $\mathbf{0}$ y $-\bar{a}(m)$ son vectores, (el resto de elementos son escalares), entonces el problema del estimado de los coeficientes por tramos se reduce a resolver la estructura matemática (B.13), a través del algoritmo que se presenta a continuación:

El algorítmo Levinson-Durbin (L-D) aplicado al tramo n $\in$ [m-N+1,m] [1][17][50][57]:

Inicialización: L=0

$\varepsilon^0(m)$ = Energía en el tramo de voz f(n;m)=S(n)W(m-n).

$= r_s(0;m)$.

Recursión: Para L = 1,2,...,M

1.- Se calcula el L-ésimo coeficiente de reflexión

$$k(L;m) = \frac{1}{\varepsilon^{L-1}(m)} \{r_s(L;m) - \sum_{i=1}^{L-1} \bar{a}^{L-1}(i;m) r_s(L-i;m)\}$$

2.- Se genera el conjunto de parámetros LP para el modelo de orden L.

$$\bar{a}^{L}(L;m) = k(L;m)$$

$$\bar{a}^{L}(i;m) = \bar{a}^{L-1}(i;m) - k(L;m)\, \bar{a}^{L-1}(L-i;m), \qquad i=1,2,....,L-1$$

3.- Se calcula la energía del error asociado con la solución de orden L.

$$\varepsilon^{L}(m) = \varepsilon^{L-1}(m)\{1 - [k(L;m)]^2\}$$

4.- Ir al paso 1 con L = L+1 siempre que L < M.

Propiedades del algorítmo L-D

Hay algunas propiedades que presenta esta recursión que se deben considerar:

1.- La secuencia de la energía promedio del error es no incremental:

$$0 \leq \varepsilon^{1}(m) \leq \varepsilon^{1-1}(m) \leq .... \leq \varepsilon^{0}(m)$$

2.- Los parámetros k se llaman coeficientes de reflexión debido a su estrecha relación con los coeficientes de reflexión de uno de tantos modelos que existen para el sistema de producción de voz, como son los modelos de tubos acústicos analógicos del tracto bucal y son de la forma:

$|k(L;m)| \leq 1$ para todo L.

Estos coeficientes juegan un rol importante en el análisis y aplicaciones de codificación de la voz y sirven como una representación paramétrica alternativa, a los coeficientes LP obtenidos por autocorrelación.

Así como los coeficientes LP se pueden obtener a partir de los coeficientes de reflexión, éstos se pueden también obtener a partir de los LP.

Hay otro conjunto de parámetros que se pueden derivar de los coeficientes de reflexión y visceversa como son los parámetros "log area ratio (LAR)" que se identifican como g(L;m), los parámetros de "seno inverso (IS)" que se identifican como G(L;m) y los parámetros cepstrales que se identifican como el CEPSTRUM [1][39][57]:

$$g(L;m) = \frac{1}{2}\log\frac{1+k(L;m)}{1-k(L;m)} = \text{Tanh}^{-1}k(L;m), \qquad L = 1,2,...,M \qquad \text{(B.14)}$$

$$G(L;m) = \frac{2}{\pi}\ \text{Sen}^{-1}k(L;m), \qquad L=1,2,...,M \qquad \text{(B.15)}$$

A los parámetros cepstrales se le dedica buena parte de lo que resta del anexo, por lo que no dan detalles en esta sección.

En algunos casos se usan los parámetros LAR e IS en lugar de los k(L;m), debido a que cuando los coeficientes de reflexión tienen una magnitud cercana a la unidad, el resultado es muy sensible a errores de cuantificación. Con la transformación a estos otros parámetros se comprime la amplitud de manera que se disminuye esa sensibilidad [1].

B.3. PRE-ÉNFASIS DE LA FORMA DE ONDA DE LA VOZ

En aplicaciones prácticas del análisis LP, es frecuente aplicar un filtro que incrementa la energía relativa del espectro de alta frecuencia antes de calcular los parámetros de predicción. Típicamente el filtro es $P(z) = 1- \mu Z^{-1}$ con $\mu \approx 1$.

Este filtro es idéntico en forma al filtro usado para modelar la radiación labial e introduce un cero cerca de w = 0. El filtro se emplea debido a que la componente de fase mínima de la señal glotal puede modelarse por un filtro simple de dos polos reales ubicados cerca de Z = 1. Además la radiación labial característica, con su cero cerca de Z = 1, tiende a cancelar los

efectos espectrales de uno de los polos glotales. Al introducir un segundo cero cerca de Z = 1, las contribuciones espectrales de la laringe y los labios se eliminan y el análisis se orienta a la búsqueda de los parámetros que corresponden al tracto bucal solamente.

El resultado del pre-énfasis es un espectro LP o un filtro libre de efectos glotales y de radiación labial (aunque hay que tener presente que estamos frente a un modelo analítico que obviamente es una simplificación de un modelo físico muy complejo). En todo caso, este procedimiento realza la influencia de los formantes más altos en el tracto bucal sobre la señal de salida s(n) [1][50].

B.4. EL ANALISIS CEPSTRAL

De acuerdo al modelo de producción tratado, la señal de voz está compuesta de una secuencia de excitación convolucionada con la respuesta impulsiva del modelo del tracto bucal. Por lo que es deseable para varios propósitos, eliminar una de las componentes tal que la otra se pueda examinar, codificar, modelar y utilizar por medio de algoritmos de reconocimiento.

En el campo de la ingeniería se sabe que eliminar una señal que se encuentra combinada con otras, es un problema bastante difícil, sin embargo, cuando su combinación es lineal hay técnicas que hacen un buen tratamiento. La herramienta por excelencia para realizar este tratamiento, que consiste en realizar en el dominio frecuencial una operación lineal sobre señales combinadas linealmente a trozos, es la Transformada de Fourier.

La Transformada de Fourier da lugar a que el espectro permita examinar las componentes individuales de las señales, debido a que éstas ocurren en lugares diferentes del espectro frecuencial, es decir, el espectro es la representación de la señal compuesta, en la cual se logra "la separación" de las partes componentes y por lo tanto se puede derivar información asociada a esas componentes.

Por ejemplo, si el objetivo fuese obtener algunas propiedades del ruido de una señal, se podría derivar información a partir del espectro. Si el objetivo fuese remover el ruido de la señal, se podría diseñar un filtro paso bajo para remover la componente indeseada de alta frecuencia y

el resultado se transformaría de nuevo al dominio del tiempo. Cada una de las operaciones realizadas para producir este resultado filtrado es un operador lineal, por lo tanto la operación total es lineal.

En lo que respecta a la señal de voz, la convolución de la secuencia de excitación con la respuesta impulsiva del tracto bucal, $s(n) = e(n) * \Theta(n)$, corresponde a una combinación no lineal, por lo que la Transformada de Fourier no proporciona mucha ayuda en forma directa. Por esa razón es que se utiliza la técnica cepstral, como una derivación de esa técnica espectral, donde al igual que el espectro, el cepstro o cepstrum representa una transformación sobre la señal de la voz con dos propiedades importantes:

1.- Las señales componentes más representativas aparecen separadas en el cepstro.
2.- Las señales componentes más representativas aparecen combinadas linealmente en el cepstro.

Como el objetivo es obtener algunas propiedades de las señales componentes, el cepstro juega un papel importante para obtener información necesaria en el procesamiento de la voz. Además, como también se tiene como objetivo eliminar alguna de las señales componentes (la excitación), se hace énfasis en el filtraje paso bajo.

Debido a que las representaciones de las señales componentes están combinadas linealmente en el cepstro, se usan filtros lineales para eliminar la componente cepstral no deseada y la componente restante se somete a la inversa de la transformación que produjo al cepstro inicial. La operación total, llamémosla μ, seguirá una suerte de principio de "superposición generalizado" [1][39].

Donde :

$$\mu\{S(n)\} = \mu\{e(n) * \Theta(\mathbf{n})\} = \mu\{e(n)\} * \mu\{\Theta(\mathbf{n})\} \qquad \text{(B.17)}$$

e(n): Secuencia de excitación.

Θ(n) : Respuesta impulsiva del tracto bucal.

El análisis cepstral es un caso especial dentro de una clase general de métodos conocida como procesamiento de señales homomórficas [1].

El cálculo del cepstro se explica siguiendo la figura B.4., es decir, de la señal de voz se calcula la magnitud espectral, luego se le aplica el logaritmo natural (puede ser cualquier base) y posteriormente al logaritmo de la magnitud espectral se le aplica la transformada inversa de Fourier.

Este dominio, que no es el dominio de la frecuencia sino una transformación de éste, se llama dominio de la cuefrencia y el cepstro o cepstrum hace el rol que tiene un espectro en el dominio de frecuencia, de la misma manera, las armónicas cuefrenciales se llaman ratmónicas en este nuevo dominio [1][8][39].

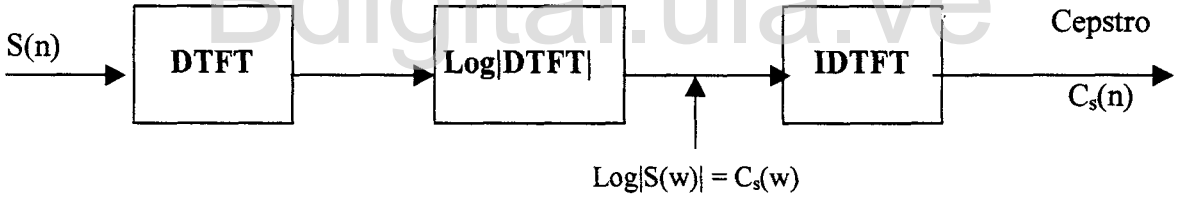


Figura B.4. Proceso de obtención de los parámetros cepstrales.

El propósito del cepstrum es separar los dos pedazos convolucionados de la voz, e(n) y Θ (n), en dos componentes aditivos y luego analizar esos componentes con análisis espectral (cepstral).

$$C_s(n) = C_e(n) + C_\Theta(n) \qquad (B.18)$$

B.4.1. Enfoque intuitivo del cepstrum

Debido a que $C_e(n)$ y $C_\Theta(n)$ están bien separadas en el dominio de la cuefrencia, se puede usar el cepstrum para eliminar, $C_e(w)$ de $C_\Theta(w)$ (hacer un filtraje en este nuevo dominio). Este

proceso de filtraje se llama liftering y en este caso se usa un "lifter tiempo bajo", análogo a un filtro paso bajo en el dominio de la frecuencia. La salida de este proceso en el dominio de la cuefrencia es un cepstrum, digamos $C_y(n) \cong C\Theta(n)$, de manera que si se desea usar $C\Theta(n)$ para obtener un estimado de $\Theta(n)$, hay que salir del dominio de la cuefrencia, invirtiendo la operación, aplicando la DTFT al cepstrum. El liftering es un proceso útil y significativo para obtener un estimado del logaritmo del espectro de cualquiera de los componentes separados. Sin embargo, si se quiere retornar al dominio de tiempo original con un estimado de la señal separada, el cepstro falla debido a que la operación de linealización no es invertible [1].

En la práctica, el cepstrum se calcula por tramos de voz de la manera siguiente:

Si, f(n;m) = S(n)W(m-n), es un tramo de longitud N que termina en el instante m, entonces la siguiente expresión es un estimador cepstral para ese tramo:

$$C_s(n;m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \{\log\left|\sum_{L=-\infty}^{\infty} f(L;m)e^{-jwL}\right|\}e^{jwn}dw$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \{\log\left|\sum_{L=m-N+1}^{m} f(l;m)e^{-jwL}\right|\}e^{jwn}dw \text{ para } n=0,1,.... \qquad (B.19)$$

Este análisis se puede observar en la figura B.5.

El cepstrum igual que como sucede con el análisis LP hace uso de la magnitud espectral solamente y no de la fase.

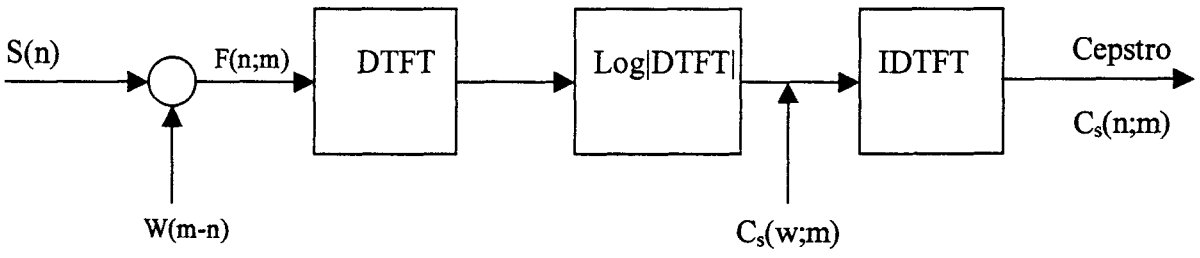


Figura B.5. Proceso de obtención de los parámetros cepstrales por tramos

B.4.2. La conversión de los parámetros LP a parámetros cepstrales

Aunque el análisis cepstral se deriva directamente del análisis espectral, una de las aplicaciones más importantes del cepstrum en el procesamiento contemporáneo de la voz es la representación de un modelo LP por parámetros cepstrales.

Los parámetros LP constituyen una representación espectral muy útil de la voz, y representan una versión del espectro que resulta de separar la excitación de la respuesta impulsiva del tracto bucal, sin embargo, el análisis LP no puede separar las características del tracto bucal de la dinámica glotal, por lo que no toma en cuenta las características laringeales que varían de persona a persona e incluso en pronunciaciones de la misma persona, lo que hace que los parámetros LP degraden el desempeño de los reconocedores, sobre todo en aquellos cuyo funcionamiento es independiente del locutor [1].

Por esa razón, en años recientes, la representación paramétrica más utilizada está constituida por el cepstro, debido a que los ingenieros del habla han encontrado mejores resultados haciendo uso de esta técnica e incluso, se utiliza una técnica que deriva los coeficientes cepstrales a partir de los coeficientes LP.

También hay que destacar que se utiliza en la actualidad, una versión del cepstro llamada MEL cepstrum, que se basa en que la percepción auditiva humana de la frecuencia no se percibe de manera lineal, es decir, se escalan los parámetros cepstrales a través de la escala de frecuencias MEL [1][4][50], se les identifica como coeficientes MFCC (MEL CEPSTRUM).

Adicionalmente, a este tipo de transformación del cepstrum, se acostumbra obtener un conjunto de parámetros que representan en forma intuitiva los cambios que sufre el espectro desde algunos tramos previamente tratados hasta el actual y desde el actual hasta otros tramos tratados posteriormente. Se habla entonces de coeficientes constituidos por la primera y segunda derivada del cepstrum [48][50][59].

5.4.3. Resúmen del procedimiento para parametrizar la voz a través de los parámetros LP y el cepstrum

1.- Realizar el pre-enfasis de la señal:

$$H(z) = 1 - a\,z^{-1} \qquad 0.9 \le a \le 1$$

$$\bar{S}(n) = S(n) - aS(n-1)$$

2.- Se divide la secuencia $\bar{S}(n)$ en segmentos de N muestras, separando los segmentos por M muestras.

Se trabaja con L tramos donde x_l es el l-ésimo tramo y $x_l(n) = S(Ml + n)$ es la n-ésima muestra del tramo l. Con n = 0,1,2,...,N-1 y l = 0,1,...,L-1.

3.- Se enventana cada tramo para minimizar la discontinuidad al inicio y al final, a través de una ventana Hamming como sigue:

$$x_l(n) = x_l(n).w(n)$$

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$$

4.- Se obtienen p+1 autocorrelaciones por cada tramo (el método de la autocorrelación es el que se usa con mayor frecuencia):

$$r_l(m) = \sum_{n=0}^{N-1-m} x_l(n)\,x_l(n+m) \qquad m = 0,1,...,p$$

5.- Cada tramo de p+1 autocorrelaciones se convierte a un conjunto de parámetros (coeficientes LP, coeficientes de reflexión, coeficientes Log area ratio, cepstrales, otra transformación, etc).

Se recurre al algoritmo Levinson-Durbin para obtener los parámetros LP de cada tramo partiendo de las autocorrelaciones de ese tramo.

6.- Luego que se obtienen los parámetros LP, se puede obtener a partir de éstos cualquier otra transformación como los parámetros LAR:

$$g_m = \log\left(\frac{1-k_m}{1-k_m}\right)$$

ó los parámetros cepstrales:

$$c_0 = \ln \sigma^2$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)\cdot c_k \cdot a_{(m-k)} \qquad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)\cdot c_k \cdot a_{(m-k)} \qquad M > p$$

σ : es la ganancia del modelo estimado.

$$\sigma = \sqrt{\frac{\varepsilon^0}{\varepsilon^p}}$$

$$\sigma^2 = 1 - \frac{1}{r(0;m)}\sum_{k=1}^{p} a(k)r(k)$$

Se calculan normalmente Q > p coeficientes cepstrales, con Q $\cong \frac{3}{2}$p

ANEXO C

INSTRUMENTOS DE TEXTO QUE SE UTILIZARON EN LA RECOLECCIÓN DE VOZ VENEZOLANA

Se generaron y se distribuyeron 2000 hojas diferentes.

A continuación se muestra el contenido general de las hojas de texto:

¿Hace la llamada desde una cabina telefónica?.....

¿Cuál es su fecha de nacimiento?.....

Hora de la llamada.....

¿En qué ciudad estudió el Bachillerato?

¿Ha nacido en Venezuela?

Lea el número de identificación número por número

texto 1	Oración
texto 2	Apellido
texto 3	Deletrear el apellido
texto 4	Oración
texto 5	Hora
texto 6	Palabra clave
texto 7	cantidad monetaria
texto 8	Oración
texto 9	Número de tarjeta de crédito
texto 10	Oración con palabras claves
texto 11	Ciudad
texto 12	Deletrear el nombre de la ciudad

texto 13	Número
texto 14	Oración
texto 15	Fecha
texto 16	Palabra
texto 17	Deletrear una palabra
texto 18	Oración
texto 19	Número telefónico
texto 20	Nombre de Agencia/compañía
texto 21	Palabra
texto 22	Oración
texto 23	Fecha relativa
texto 24	Palabra clave
texto 25	Palabra
texto 26	Oración
texto 27	Palabra
texto 28	Dígito
texto 29	Palabra clave
texto 30	Oración
texto 31	Número de 6 dígitos
texto 32	Palabra clave
texto 33	Oración
texto 34	Dígitos
texto 35	Palabra clave
texto 36	Nombre
texto 37	Palabra clave
texto 38	Oración adicional

EJEMPLO DE LA HOJA DE INSTRUCCIONES Y DEL TEXTO ENTREGADO A UN LOCUTOR

UNIVERSIDAD DE LOS ANDES

COLABORACION EN UN PROYECTO DE INVESTIGACION

INSTRUCCIONES

Lea cuidadosamente estas instrucciones **antes de llamar** al sistema de grabación.

- Esta hoja contiene las preguntas que le formulará el sistema de grabación y los textos que debe leer. **Lea en voz alta los textos**, especialmente si hay alguno difícil de pronunciar.
- Es conveniente que realice la llamada en un momento en que ningún aparato (lavadora, etc.) pueda producir un ruido de fondo; **apague su radio, televisor o equipo de sonido**. Evite mantener una conversación con alguien más, mientras nos esté telefoneando.
- Asegúrese de conocer **la fecha y la hora** en que realiza la llamada así como su número de teléfono. Puede serle útil escribir estos datos antes de llamar.
- Cuando realice la llamada, **hable normalmente,** con su tono y velocidad habituales. **Espere hasta oír un tono antes de responder** a las preguntas o leer cada texto.
- Algunos textos que contienen números conviene leerlos dígito a dígito (por ejemplo: 171: uno, siete, uno) El sistema se lo indicará oportunamente.
- **No se preocupe si se equivoca**, siga adelante y nosotros lo corregiremos más adelante.
- Si su voz no se escucha adecuadamente, el sistema repetirá la última pregunta o solicitará que lea otra vez la última frase. Si el sistema detecta fallos en la comunicación debidos a problemas en la línea, le informará que hay un problema y cancelará la llamada. Por favor, trate de realizar la llamada un poco más tarde.

Muchas gracias por su colaboración.

AHORA, POR FAVOR, LLAME AL NUMERO

XXXX XX XX XX

Un sistema automático le responderá y le facilitará las instrucciones a seguir .

(A continuación se detalla el mensaje que el sistema automático le dará)

Gracias por llamar a nuestro sistema de grabación automático. La sesión empieza con unas preguntas. Por favor hable en la forma habitual, después de oír el tono.

¿Hace la llamada desde una cabina telefónica?.....

¿Cuál es su fecha de nacimiento?.....

Hora de la llamada.....

¿En qué ciudad estudió el Bachillerato?

¿Ha nacido en Venezuela?

Gracias, ahora gire la hoja

Lea el número de identificación dígito por dígito **000000**

A continuación se le pedirá que lea cada uno de los textos de la columna de la derecha

texto 1	El general señaló que no pudieron responder a los disparos
texto 2	Robles
texto 3	R O B L E S
texto 4	Llevará a cabo una reducción progresiva de los tipos de interés
texto 5	son exactamente las tres y diecisiete minutos de la tarde
texto 6	Guarda
texto 7	601.704 bolívares
texto 8	Dirigiéndose a una audiencia superior y más alta.
texto 9	1604 8388 6929 7557
texto 10	Vuelve a llamar al mismo número
texto 11	Mérida
texto 12	M E R I D A
texto 13	115.502
texto 14	Tiró la locomotora y advirtió el riesgo que corrían
texto 15	Lunes, trece de mayo de 1928

texto 16	Hinojosa
texto 17	P Z U H CH Q Y
texto 18	Va a ejercer su derecho a veto respecto del ingreso de algunos países
texto 19	051-330000
texto 20	TROC
texto 21	Bezaar
texto 22	Al fin y al cabo no daba un duro por su vida.
texto 23	el próximo mes de marzo
texto 24	Grabar
texto 25	Potingue
texto 26	Al mismo tiempo se ha puesto en contacto con el Departamento de Justicia
texto 27	Substraendo
texto 28	8
texto 29	Volver a llamar
texto 30	He pasado del día agridulce a la euforia iconoclasta.
texto 31	1 5 4 4 0 0
texto 32	Añadir
texto 33	Tomaron conciencia de que merecía la pena ser indio.
texto 34	7, 5, 9, 0, 8, 1, 4, 6, 2, 3
texto 35	Borrar
texto 36	Juan Anglada
texto 37	Reproducir
texto 38	El merengue islandés le embarró un mechón

Gracias por su colaboración. Se ha realizado una grabación completa. Adiós.

Tabla C.1. El conjunto de las frases de aplicación.

Venezuela es un gran país	finalizar el programa
Quiero una conexión con Venezuela	pasar al mensaje siguiente
deseo ir al siguiente programa	finalizar la multiconferencia
en español me expresaré mejor	marcar el siguiente número
Quiero insertar un nombre	deseo volver a marcar a la recepción
intenta cambiar el idioma	quiero borrar el mensaje anterior
repetir el mensaje número ocho	deseo desconectar inmediatamente

insertar otro número en las memorias	parar la reproducción de mensajes
insertar un nuevo programa	borrar ese nombre de la agenda
la tecla asterisco no puede ser marcada	en español, por favor
quiero acceder al menú de ayuda	finalizar con la introducción de datos
has de volver a marcar hasta que consigas la llamada	para leer el mensaje marque asterisco
por favor utilice el asterisco	borrar el número de la memoria
por favor, marcar el número de mi casa	desconectar el teléfono de la línea
desconectar todos los teléfonos	después de grabarlo se debe enviar a la extensión ciento doce
Transferir la llamada a la secretaria	marcar el cero cero tres.
me gusta vivir en Venezuela	ir al final del mensaje
continuar con la ayuda	llamar al servicio de información
insertar nuevo teléfono	viajará desde Venezuela
he de volver a marcar al mismo número	cambiar el número de mi esposa
muéstrame la lista de opciones	si no contesta cancelar la orden
ahora continuar con la llamada	reproducir todos los mensajes recibidos
enviar aviso	después de esto, finalizar la sesión
reproduce el directorio principal	quiero enviar la grabación al cero dos ocho
utilizar el siguiente número de teléfono	guardar un nuevo teléfono
Estoy en Venezuela	transferir la siguiente llamada
Transferir mis llamadas a la oficina	se debe guardar este mensaje
¿podría hablar en español?	la operadora lo atenderá en un momento
Sería mejor si pudiera hablar en español	grabar el número en la cuarta posición
el anterior, por favor	mi idioma es el español
finalizar la operación en curso	ahora grabar un mensaje para mi esposa
Favor desconectar el teléfono	continuar después de la transferencia
quiero parar la grabación	al finalizar repetir el comando
finalizar con el servicio de agenda	es para borrar el mensaje
ir al menú anterior	deseo consultar el siguiente mensaje
Continuar con la entrada de datos	favor, llamar a mi secretaria
cambiar el programa anterior	cambiar la clave de acceso
se debe desconectar la línea	grabar el final de la llamada
repetir la información del directorio	cancelar la grabación del mensaje
lista todo el directorio	reproducir el contenido del menú
llamar a la extensión seis cuatro cinco siete	llamar a los bomberos de la ciudad
marcar el teléfono de la operadora	marcar el teléfono de la operadora

ayuda sobre el servicio.	desea comunicarse con la operadora
borrar la actividad ocho	solicita parar la llamada
guardar la lista	es la tecla para grabar el mensaje
Continuar la reproducción del mensaje	mi número telefónico es el cuatro dos uno tres
ahora muestra las opciones de directorio	ir a la quinta posición de la lista
enviar el mensaje recibido al otro teléfono	guardar el teléfono de Alberto Gómez
repetir el programa del día	quisiera obtener ayuda sobre esta opción
¿podría oír el mensaje de ayuda?	por favor, repetir la llamada
dentro de diez minutos llamar otra vez	grabar el siguiente mensaje
hunda la tecla asterisco después de la señal	si no contestan, volver a marcar
borrar el final de la cinta	muestra el directorio
por favor, transferir todas las llamadas a recepción	no tengo número de acceso
borrar el mensaje anterior	avanzar hasta el mensaje siguiente
Siguiente opción	la operadora se encuentra ocupada
leer el final del mensaje	por favor, cancelar la llamada
añade una entrada al directorio	marcar el teléfono de la policía
Favor comunicarme con la operadora	agregar un nuevo número a la lista
grabar un mensaje general	reproducir la información de deportes
ve al directorio, por favor	la llamada ha llegado al final
Ahora repetir el proceso de enviar	quiero cambiar a la memoria cinco
quiero introducir el número de acceso	transferir la llamada a la central telefónica
Parar la operación en curso	cancelar el último número memorizado
Parar el servicio de mensajería	reproducir el mensaje grabado
Ahora cancelar la memorización de los nombres	cambiar la configuración original
continuar reproduciendo el siguiente mensaje	insertar una nueva entrada en la agenda
el número del producto es el dos uno dos	lista las opciones
¿cuáles son las opciones del servicio?	debe hundir la tecla asterisco
guardar los mensajes que no he escuchado	a continuación reproducir la información de carreteras
quiero oír el mensaje anterior	repetir el mensaje anterior
necesito ayuda sobre la opción dos	repite las opciones posibles
insertar otra opción en el menú	dime las opciones actuales
ahora guardar los mensajes recibidos	enviar el mensaje siete a los jefes de sección
¿qué opciones están disponibles?	a la hora fijada hay que volver a marcar

ANEXO D

MUESTRAS DE RESULTADOS EN EL RECONOCIMIENTO DEL HABLA VENEZOLANA

SALIDADIGITOSENT4

```
------------------------ Sentence Scores ------------------------
====================== HTK Results Analysis =====================
  Date: Thu Apr 25 15:22:20 2002
  Ref : /home/luciano/dialectos/listas/totaldigitosx.mlf
  Rec : /home/luciano/dialectos/reconocidos/digitos/MLFsalidadigitos4
------------------------- File Results --------------------------

                                    .
                                    .
                                    .
```



```
a40067c1.rec:  100.00( 75.00)  [H=   4, D=  0, S=  0, I=  1, N=  4]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40067c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40067c1.rec
 LAB: dos uno uno dos
 REC: dos uno uno dos seis
a40067c4.rec:  100.00(100.00)  [H=   4, D=  0, S=  0, I=  0, N=  4]
a40176c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40176c3.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40054c1.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40054c2.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40054c3.rec:  100.00(100.00)  [H=   8, D=  0, S=  0, I=  0, N=  8]
a40054c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40146c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40146c3.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40146c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40184c1.rec:  100.00(100.00)  [H=   4, D=  0, S=  0, I=  0, N=  4]
a40184c4.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40285c1.rec:   83.33( 83.33)  [H=   5, D=  0, S=  1, I=  0, N=  6]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40285c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40285c1.rec
 LAB: tres tres tres tres cinco seis
 REC: tres seis tres tres cinco seis
a40407c1.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40407c2.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40070c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40070c3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40070c4.rec:  100.00( 83.33)  [H=   6, D=  0, S=  0, I=  1, N=  6]
```

```
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40070c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40070c4.rec
 LAB: seis       seis ocho cero dos seis
 REC: seis siete seis ocho cero dos seis
a40074c1.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40087c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40087c3.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40087c4.rec:  100.00( 83.33)  [H=   6, D=  0, S=  0, I=  1, N=  6]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40087c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40087c4.rec
 LAB: ocho     dos ocho tres cinco cinco
 REC: ocho dos dos ocho tres cinco cinco
a40107c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40107c3.rec:   87.50( 87.50)  [H=   7, D=  0, S=  1, I=  0, N=  8]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40107c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40107c3.rec
 LAB: dos siete siete seis ocho seis seis cinco
 REC: dos siete siete seis ocho seis seis siete
a40107c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40110c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40110c3.rec:  100.00(100.00)  [H=   4, D=  0, S=  0, I=  0, N=  4]
a40110c4.rec:   83.33( 83.33)  [H=   5, D=  0, S=  1, I=  0, N=  6]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40110c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40110c4.rec
 LAB: cuatro cero seis siete siete cinco
 REC: cuatro cero seis siete siete siete
a40307c1.rec:  100.00(100.00)  [H=   4, D=  0, S=  0, I=  0, N=  4]
a40307c2.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40307c3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40307c4.rec:  100.00( 75.00)  [H=   4, D=  0, S=  0, I=  1, N=  4]
Aligned transcription:
/home/luciano/etiquetas/digitos/total/total2/a40307c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40307c4.rec
 LAB: tres cuatro cuatro siete
 REC: tres cuatro cuatro siete tres
------------------------ Overall Results --------------------------
SENT: %Correct=75.35 [H=321, S=105, N=426]
WORD: %Corr=99.33, Acc=94.77 [H=2508, D=7, S=10, I=115, N=2525]
==========================================================================
```

SALIDADIGITOSTEST7

```
------------------------ Sentence Scores -----------------------
====================== HTK Results Analysis =====================
  Date: Fri Apr 26 09:40:43 2002
  Ref : /home/luciano/dialectos/listas/digitostestcasa1.mlf
  Rec : /home/luciano/dialectos/reconocidos/digitos/MLFsalidadigitos6
------------------------- File Results --------------------------

                                    .
                                    .
                                    .

a40096c1.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40096c3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40096c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40097c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40097c3.rec:  100.00( 87.50)  [H=   8, D=  0, S=  0, I=  1, N=  8]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40097c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40097c3.rec
 LAB: uno nueve nueve cinco cinco cero nueve      uno
 REC: uno nueve nueve cinco cinco cero nueve cero uno
a40097c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40148c1.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40148c2.rec:  100.00( 87.50)  [H=   8, D=  0, S=  0, I=  1, N=  8]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40148c2.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40148c2.rec
 LAB: cero cero siete cuatro nueve  siete dos cero
 REC: cero cero siete cuatro nueve tres siete dos cero
a40148c3.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40148c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40247c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40247c3.rec:  100.00(100.00)  [H=   8, D=  0, S=  0, I=  0, N=  8]
a40247c4.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40249c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40249c3.rec:   85.71( 85.71)  [H=   6, D=  0, S=  1, I=  0, N=  7]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40249c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40249c3.rec
 LAB: cero tres cero siete cinco nueve cero
 REC: cero tres cero siete cinco nueve tres
a40249c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40251c1.rec:  100.00(100.00)  [H=   2, D=  0, S=  0, I=  0, N=  2]
a40251c2.rec:  100.00( 80.00)  [H=   5, D=  0, S=  0, I=  1, N=  5]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40251c2.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40251c2.rec
 LAB: cero cero dos      ocho ocho
 REC: cero cero dos cero ocho ocho
a40251c3.rec:  100.00(100.00)  [H=   4, D=  0, S=  0, I=  0, N=  4]
a40251c4.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40252c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40252c2.rec:  100.00( 80.00)  [H=   5, D=  0, S=  0, I=  1, N=  5]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40252c2.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40252c2.rec
```

```
 LAB: cero cero dos      ocho ocho
 REC: cero cero dos cero ocho ocho
a40252c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40257c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40257c4.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40258c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40258c2.rec:   83.33( 83.33)  [H=   5, D=  1, S=  0, I=  0, N=  6]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40258c2.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40258c2.rec
 LAB: cero cero dos ocho ocho cinco
 REC: cero cero     ocho ocho cinco
a40282c1.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40282c2.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40282c3.rec:  100.00( 88.89)  [H=   9, D=  0, S=  0, I=  1, N=  9]
Aligned transcription:
/home/luciano/etiquetas/digitos/test/test2/a40282c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40282c3.rec
 LAB: cinco tres siete uno nueve tres dos nueve ocho
 REC: cinco tres siete uno nueve tres dos nueve ocho uno
----------------------- Overall Results -------------------------
SENT: %Correct=64.29 [H=81, S=45, N=126]
WORD: %Corr=98.47, Acc=92.09 [H=710, D=1, S=10, I=46, N=721]
=============================================================
```


FONOSMUJERES

Nro. de Fono	Símbolo	Patrones de entrenamiento
1	"B"	15
2	"D"	17
3	"G"	5
4	"M"	4
5	"N"	315
6	"R"	4
7	"a"	254
8	"b"	319
9	"c"	55
10	"d"	374
11	"e"	1345
12	"f"	15
13	"g"	31
14	"h"	49
15	"i"	391
16	"j"	337
17	"k"	90
18	"l"	229
19	"m"	228
20	"n"	263
21	"o"	642
22	"p"	22
23	"r"	281

24	"s"	597
25	"t"	481
26	"u"	81
27	"w"	98
28	"y"	27
29	"sil"	419

DICCIONARIOFECHAS

abril a b r i l
agosto a g o s t o
agosto a g o h t o
año a M o
catorce k a t o r s e
cero s e r o
cinco s i N k o
cincuenta s i N k w e N t a
cincuenta s i N k w e N t e
cuarenta k w a r e N t a
cuarenta k w a r e N t e
cuatro k w a t r o
de d e
de e
de D e
de d i
del d e l
del D e l
diciembre d i s j e m b r e
diciembre D i s j e m b r e
diciembre d i s j e m B r e
diciembre D i s j e m B r e
diecinueve d j e s i n w e b e
diecinueve D j e s i n w e b e
dieciocho d j e s i o c o
dieciocho D j e s i o c o
dieciseis d j e s i s e j s
dieciseis d j e s i s e j
dieciseis D j e s i s e j s
dieciseis D j e s i s e j
diecisiete d j e s i s j e t e
diecisiete D j e s i s j e t e
diez d j e s
diez d j e
diez D j e s
diez D j e
diez d j e h
diez D j e h
doce d o s e
doce D o s e
domingo d o m i N G o
domingo D o m i N G o

dos d o s
dos d o
dos D o s
dos D o
dos d o h
dos D o h
el e l
en e n
enero e n e r o
febrero f e b r e r o
jueves h w e b e s
jueves h w e b e
jueves h w e b e h
julio h u l j o
junio h u n j o
lunes l u n e s
lunes l u n e h
lunes l u n e
martes m a r t e s
martes m a r t e
martes m a r t e h
marzo m a r s o
mayo m a y o
miercoles m j e r k o l e s
miercoles m j e r k o l e
miercoles m j e r k o l e h
mil m i l
mil m i
novecientos n o b e s j e N t o s
novecientos n o b e s j e N t o
novecientos n o b e s j e N t o h
noventa n o b e N t a
noventa n o b e N t i
noviembre n o b j e m b r e
noviembre n o b j e m B r e
nueve n w e b e
ochenta o c e N t a
ochenta o c e N t e
ochenta o c e N t i
ocho o c o
octubre o t u b r e
once o n s e
quince k i n s e
sabado s a b a d o
seis s e j s
seis s e j
septiembre s e t j e m b r e
septiembre s e t j e m B r e
sesenta s e s e N t a
sesenta s e s e N t e
setenta s e t e N t a
setenta s e t e N t e
setenta s e t e N t i
siete s j e t e
trece t r e s e
treinta t r e j N t a
tres t r e s

tres t r e
tres t r e h
uno u n o
veinte b e j N t e
veinte B e j N t e
veinte b e j N t i
veinticinco B e j N t i s i N k o
veinticinco b e j N t i s i N k o
veinticuatro b e j N t i k w a t r o
veinticuatro B e j N t i k w a t r o
veintidos b e j N t i d o s
veintidos B e j N t i d o
veintidos b e j N t i d o h
veintidos B e j N t i d o h
veintidos B e j N t i d o s
veintidos b e j N t i d o
veintidos b e j N t i o s
veintidos B e j N t i o
veintidos B e j N t i o h
veintinueve b e j N t i n w e b e
veintinueve B e j N t i n w e b e
veintiocho b e j N t i o c o
veintiocho B e j N t i o c o
veintiseis b e j N t i s e j s
veintiseis b e j N t i s e j
veintiseis B e j N t i s e j s
veintiseis b e j N t i s e j h
veintiseis B e j N t i s e j h
veintiseis B e j N t i s e j
veintisiete b e j N t i s j e t e
veintisiete B e j N t i s j e t e
veintitres b e j N t i t r e s
veintitres b e j N t i t r e
veintitres B e j N t i t r e s
veintitres b e j N t i t r e h
veintitres B e j N t i t r e h
veintitres B e j N t i t r e
veintiuno b e j N t i u n o
veintiuno B e j N t i u n o
viernes b j e r n e s
viernes b j e r n e
viernes B j e r n e s
viernes b j e r n e h
viernes B j e r n e h
viernes B j e r n e
y i
y y
sil sil
sp sp

GRAMMUJERES3

```
$num1 = cero | uno | dos | tres | cuatro | cinco | seis | siete | ocho |
nueve;
$num2 = diez | once | doce | trece | catorce | quince | dieciseis |
diecisiete | dieciocho | diecinueve | veinte | veintiuno | veintidos |
veintitres | veinticuatro | veinticinco | veintiseis | veintisiete |
veintiocho | veintinueve | treinta;

$num3 = cuarenta | cincuenta | sesenta | setenta | ochenta | noventa | sil;

$fin2 = y $num1;
$fin1 = $num1 | $num2 | $num3 $fin2 | $num3;

$nmes = $num1 | diez | once | doce;

$dia = lunes | martes | miercoles | jueves | viernes | sabado | domingo |
sil;

$mes = enero | febrero | marzo | abril | mayo | junio | julio | agosto |
septiembre | octubre | noviembre | diciembre;


$deldosmil = del dos mil;

$demil = de mil | del mil;

$dosmil = $deldosmil;

$ano = $demil;
$demes = $mes | de $mes;
$eldia = el $dia | $dia;
$enmes = en $mes;
$num = $num1| $num2;
$elnum = el $num | $num;

$parte4 = $eldia [sil] $num [y uno]| el $num | $num;
$parte1 = $parte4 de $mes [sil]| $enmes | $elnum $nmes| treinta y uno;
$parte2 = $ano | $dosmil;
$parte3 = $fin1;
$parte5 = $parte2 novecientos [sil]| $parte2;
$parte6 = $parte5 $parte3 | del $parte3 | $parte3;
$parte7 = $parte4 $demes;
$pasado = el $dia pasado| el mes pasado;

$proximo = el proximo mes de $mes | el proximo $dia;
( [sil] ( ($parte1 $parte5)|($parte1 $parte6)| $parte7 | $pasado | $proximo
) [sil])
```

RESULTADOSMUJERES6

```
---------------------- Sentence Scores ----------------------
====================== HTK Results Analysis ======================
  Date: Tue Feb 26 09:20:55 2002
  Ref : /home/luciano/tesisfechas/listas/palabrasmujerestest.mlf
  Rec : /home/luciano/tesisfechas/reconocidos/MLFsalida
-------------------------- File Results --------------------------

                                    .
                                    .
                                    .


a40081d2.rec:  100.00(100.00)  [H=  12, D=  0, S=  0, I=  0, N= 12]
a40081d3.rec:   85.71( 85.71)  [H=   6, D=  0, S=  1, I=  0, N=  7]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40081d3.lab vs
/home/luciano/tesisfechas/reconocidos/a40081d3.rec
 LAB: sil el proximo mes de mayo  sil
 REC: sil el proximo mes de marzo sil
a40085d1.rec:   87.50( 87.50)  [H=   7, D=  0, S=  1, I=  0, N=  8]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40085d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40085d1.rec
 LAB: sil diez      ocho del ochenta y uno sil
 REC: sil dieciocho ocho del ochenta y uno sil
a40085d2.rec:  100.00(100.00)  [H=  14, D=  0, S=  0, I=  0, N= 14]
a40107d3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40139d2.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40139d3.rec:   83.33( 66.67)  [H=   5, D=  0, S=  1, I=  1, N=  6]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40139d3.lab vs
/home/luciano/tesisfechas/reconocidos/a40139d3.rec
 LAB:    el proximo mes de mayo  sil
 REC: sil el proximo mes de marzo sil
a40146d2.rec:  100.00(100.00)  [H=  13, D=  0, S=  0, I=  0, N= 13]
a40146d3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40176d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40176d2.rec:   90.91( 81.82)  [H=  10, D=  0, S=  1, I=  1, N= 11]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40176d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40176d2.rec
 LAB: sil viernes sil veinticuatro de septiembre de mil sil   quince sil
 REC: sil viernes sil veinticuatro de septiembre de mil sil y seis   sil
a40176d3.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40184d1.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40184d2.rec:  100.00(100.00)  [H=   8, D=  0, S=  0, I=  0, N=  8]
a40184d3.rec:  100.00( 60.00)  [H=   5, D=  0, S=  0, I=  2, N=  5]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40184d3.lab vs
/home/luciano/tesisfechas/reconocidos/a40184d3.rec
 LAB:     el     tres de marzo sil
 REC: sil el sil tres de marzo sil
a40257d2.rec:  100.00(100.00)  [H=  10, D=  0, S=  0, I=  0, N= 10]
a40285d3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40301d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
```

```
a40302d2.rec:  100.00(100.00)  [H=  13, D=  0, S=  0, I=  0, N= 13]
a40407d2.rec:   90.91( 90.91)  [H=  10, D=  0, S=  1, I=  0, N= 11]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40407d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40407d2.rec
 LAB: sil martes veinticuatro de mayo sil del dos mil cuatro sil
 REC: sil martes veinticuatro de mayo sil del dos mil dos    sil
a40433d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40433d2.rec:  100.00( 90.00)  [H=  10, D=  0, S=  0, I=  1, N= 10]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40433d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40433d2.rec
 LAB: sil    miercoles catorce de mayo del dos mil veinticuatro sil
 REC: sil el miercoles catorce de mayo del dos mil veinticuatro sil
a40476d2.rec:   90.91( 90.91)  [H=  10, D=  1, S=  0, I=  0, N= 11]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40476d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40476d2.rec
 LAB: sil domingo veintiocho de septiembre de del dos mil diecinueve sil
 REC: sil domingo veintiocho de septiembre    del dos mil diecinueve sil
a40479d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40479d2.rec:   91.67( 91.67)  [H=  11, D=  0, S=  1, I=  0, N= 12]
Aligned transcription:
/home/luciano/etiquetas/mujeres/fechas/test/palabras/a40479d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40479d2.rec
 LAB: martes sil veintiocho de febrero de mil novecientos sesenta y seis
sil
 REC: martes sil veintiocho de febrero de mil novecientos setenta y seis
sil
------------------------ Overall Results ----------------------
SENT: %Correct=60.53 [H=23, S=15, N=38]
WORD: %Corr=96.98, Acc=95.05 [H=353, D=1, S=10, I=7, N=364]
================================================================
```


FONOSTOTAL

Nro. de Fono	Símbolo	Patrones de entrenamiento
1	"B"	15
2	"D"	17
3	"G"	5
4	"M"	8
5	"N"	431
6	"R"	5
7	"a"	363
8	"b"	445
9	"c"	76
10	"d"	558
11	"e"	1884
12	"f"	21
13	"g"	40

14	"h"	75
15	"i"	541
16	"j"	479
17	"k"	134
18	"l"	332
19	"m"	335
20	"n"	371
21	"o"	929
22	"p"	31
23	"r"	400
24	"s"	869
25	"t"	674
26	"u"	125
27	"w"	140
28	"y"	41
29	"sil"	621

SALIDATOTALTEST3

```
------------------------ Sentence Scores ------------------------
====================== HTK Results Analysis =====================
  Date: Mon Mar 11 17:46:16 2002
  Ref : /home/luciano/tesisfechas/listas/palabrastotaltest3.mlf
  Rec : /home/luciano/tesisfechas/reconocidos/MLFsalidatotaltest3
------------------------ File Results --------------------------

                                    .
                                    .
                                    .


a40001d1.rec:   80.00( 70.00)  [H=   8, D=  1, S=  1, I=  1, N= 10]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40001d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40001d1.rec
 LAB: sil catorce   sil de marzo del setenta y cinco sil
 REC: sil catorce y uno de marzo     setenta y cinco sil
a40003d2.rec:  100.00(100.00)  [H=  10, D=  0, S=  0, I=  0, N= 10]
a40077d1.rec:   87.50( 87.50)  [H=   7, D=  1, S=  0, I=  0, N=  8]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40077d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40077d1.rec
 LAB: sil trece nueve del cuarenta y cuatro sil
 REC: sil trece nueve     cuarenta y cuatro sil
a40077d2.rec:   91.67( 91.67)  [H=  11, D=  0, S=  1, I=  0, N= 12]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40077d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40077d2.rec
 LAB: sil martes dieciocho de enero de mil novecientos treinta y dos sil
 REC: sil martes dieciocho de enero de mil novecientos noventa y dos sil
a40097d3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40148d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40148d2.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40251d2.rec:   90.91( 90.91)  [H=  10, D=  0, S=  1, I=  0, N= 11]
```

```
Aligned transcription: /home/luciano/etiquetas/fechastest/a40251d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40251d2.rec
 LAB: sil sabado sil cinco de enero sil de  dos mil uno
 REC: sil sabado sil cinco de enero sil del dos mil uno
a40252d1.rec:   87.50( 87.50)  [H=   7, D=  1, S=  0, I=  0, N=  8]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40252d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40252d1.rec
 LAB: sil dos dos del setenta y tres sil
 REC: sil dos dos     setenta y tres sil
a40252d2.rec:  100.00(100.00)  [H=  10, D=  0, S=  0, I=  0, N= 10]
a40270d2.rec:  100.00( 90.00)  [H=  10, D=  0, S=  0, I=  1, N= 10]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40270d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40270d2.rec
 LAB: sil    lunes ocho de abril de mil novecientos noventa sil
 REC: sil el lunes ocho de abril de mil novecientos noventa sil
a40270d3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40308d2.rec:  100.00(100.00)  [H=  12, D=  0, S=  0, I=  0, N= 12]
a40308d3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40325d1.rec:  100.00( 91.67)  [H=  12, D=  0, S=  0, I=  1, N= 12]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40325d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40325d1.rec
 LAB:     treinta y uno de marzo de mil novecientos cuarenta y cuatro sil
 REC: sil treinta y uno de marzo de mil novecientos cuarenta y cuatro sil
a40325d3.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40330d1.rec:   83.33( 83.33)  [H=   5, D=  1, S=  0, I=  0, N=  6]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40330d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40330d1.rec
 LAB: sil veintiocho doce cinco dos sil
 REC: sil veintiocho doce cinco     sil
a40060d3.rec:  100.00(100.00)  [H=   6, D=  0, S=  0, I=  0, N=  6]
a40066d2.rec:   90.91( 90.91)  [H=  10, D=  0, S=  1, I=  0, N= 11]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40066d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40066d2.rec
 LAB: sil martes seis de agosto de mil novecientos cincuenta y uno
 REC: sil martes seis de agosto de mil novecientos cincuenta y dos
a40067d2.rec:  100.00(100.00)  [H=  13, D=  0, S=  0, I=  0, N= 13]
a40070d2.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40081d2.rec:  100.00(100.00)  [H=  12, D=  0, S=  0, I=  0, N= 12]
a40081d3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40085d1.rec:   50.00(  0.00)  [H=   4, D=  1, S=  3, I=  4, N=  8]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40085d1.lab vs
/home/luciano/tesisfechas/reconocidos/a40085d1.rec
 LAB: sil diez    ocho          del         ochenta  y uno sil
 REC: sil viernes ocho de abril del dos mil cuarenta y     dos
a40085d2.rec:  100.00(100.00)  [H=  14, D=  0, S=  0, I=  0, N= 14]
a40107d3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40139d2.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40139d3.rec:  100.00(100.00)  [H=   7, D=  0, S=  0, I=  0, N=  7]
a40146d2.rec:   92.31( 92.31)  [H=  12, D=  0, S=  1, I=  0, N= 13]
Aligned transcription: /home/luciano/etiquetas/fechastest/a40146d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40146d2.rec
 LAB: sil miercoles sil dieciseis de febrero de  mil novecientos noventa y
tres sil
 REC: sil miercoles sil dieciseis de febrero del mil novecientos noventa y
tres sil

a40433d2.rec:  100.00(100.00)  [H=  10, D=  0, S=  0, I=  0, N= 10]
a40476d2.rec:   90.91( 90.91)  [H=  10, D=  1, S=  0, I=  0, N= 11]
```

```
Aligned transcription: /home/luciano/etiquetas/fechastest/a40476d2.lab vs
/home/luciano/tesisfechas/reconocidos/a40476d2.rec
 LAB: sil domingo veintiocho de septiembre de del dos mil diecinueve sil
 REC: sil domingo veintiocho de septiembre    del dos mil diecinueve sil
a40476d3.rec:  100.00(100.00)  [H=   5, D=  0, S=  0, I=  0, N=  5]
a40479d1.rec:  100.00(100.00)  [H=  11, D=  0, S=  0, I=  0, N= 11]
a40479d2.rec:  100.00(100.00)  [H=  12, D=  0, S=  0, I=  0, N= 12]
----------------------- Overall Results -------------------------
SENT: %Correct=70.31 [H=45, S=19, N=64]
WORD: %Corr=96.82, Acc=94.14 [H=578, D=7, S=12, I=16, N=597]
```

FONOSHOMBRES

Nro. de Fono	Símbolo	Patrones de entrenamiento
1	"M"	4
2	"N"	116
3	"R"	4
4	"a"	109
5	"b"	126
6	"c"	21
7	"d"	184
8	"e"	539
9	"f"	6
10	"g"	9
11	"h"	26
12	"i"	150
13	"j"	142
14	"k"	44
15	"l"	103
16	"m"	107
17	"n"	108
18	"o"	287
19	"p"	9
20	"r"	119
21	"s"	272
22	"t"	193
23	"u"	44
24	"w"	42
25	"y"	14
26	"sil"	202

ESTADISTICA10

Número	Modelo	Realizaciones
1	"acuatroh"	17
2	"aochof"	24
3	"aochoh"	11
4	"odosf"	31
5	"aunof"	25
6	"odosh"	10
7	"aunoh"	8
8	"zseisf"	21
9	"zseish"	25
10	"ocincof"	35
11	"ocincoh"	2
12	"llcincof"	32
13	"llcincoh"	17
14	"csietef"	37
15	"csieteh"	20
16	"acincof"	16
17	"acincoh"	20
18	"cochof"	27
19	"cochoh"	22
20	"sil"	1671
21	"zdosf"	29
22	"zdosh"	29
23	"zsietef"	15
24	"onuevef"	15
25	"zsieteh"	27
26	"oseisf"	26
27	"onueveh"	11
28	"oseish"	10
29	"adosf"	30
30	"adosh"	36
31	"llnuevef"	27
32	"llnueveh"	21
33	"aseisf"	22
34	"aseish"	13
35	"anuevef"	15
36	"lltresf"	32
37	"anueveh"	10
38	"lltresh"	20
39	"zcuatrof"	27
40	"zcuatroh"	15
41	"llcuatrof"	53
42	"llcuatroh"	31
43	"cunof"	35
44	"cunoh"	30
45	"ccincof"	35
46	"ccincoh"	25
47	"llcerof"	43
48	"llceroh"	33
49	"cseisf"	31
50	"cseish"	34
51	"ztresf"	32
52	"ztresh"	31
53	"ccuatrof"	34

```
 54     "ccuatroh"   25
 55     "ocuatrof"   19
 56     "ocuatroh"    8
 57       "zcerof"   25
 58       "zceroh"   34
 59      "zcincof"   32
 60      "zcincoh"   40
 61      "llochof"   32
 62      "llochoh"   15
 63       "llunof"   37
 64       "llunoh"   21
 65      "cnuevef"   40
 66      "cnueveh"   36
 67        "cdosf"   46
 68        "cdosh"   32
 69       "otresf"   33
 70       "otresh"   22
 71       "zochof"   18
 72       "zochoh"   24
 73      "osietef"   27
 74      "osieteh"   11
 75        "ounof"   21
 76       "atresf"   21
 77        "ounoh"    7
 78       "atresh"   13
 79      "znuevef"   27
 80      "znueveh"   24
 81     "llsietef"   41
 82     "llsieteh"   23
 83       "ocerof"   26
 84       "oceroh"    9
 85      "asietef"   17
 86      "asieteh"    7
 87       "acerof"   33
 88       "aceroh"   19
 89       "lldosf"   44
 90       "lldosh"   24
 91      "llseisf"   39
 92      "llseish"   25
 93       "ctresf"   35
 94       "ctresh"   45
 95        "zunof"   19
 96        "zunoh"   22
 97       "oochof"   36
 98       "oochoh"   15
 99       "ccerof"   38
100       "cceroh"   27
101     "acuatrof"   14
```

GRAMATICATOTAL1

```
$digito1 = zceroh | zunoh | zdosh | ztresh | zcuatroh | zcincoh | zseish |
zsieteh | zochoh | znueveh |sil;
$digito2 = zcerof | zunof | zdosf | ztresf | zcuatrof | zcincof | zseisf |
zsietef | zochof | znuevef |sil;
$digito3 = aceroh | aunoh | adosh | atresh | acuatroh | acincoh | aseish |
asieteh | aochoh | anueveh |sil;
$digito4 = acerof | aunof | adosf | atresf | acuatrof | acincof | aseisf |
asietef | aochof |  anuevef |sil;

$digito5 = cceroh | cunoh | cdosh | ctresh | ccuatroh | ccincoh | cseish |
csieteh | cochoh | cnueveh |sil;
$digito6 = ccerof | cunof | cdosf | ctresf | ccuatrof | ccincof | cseisf |
csietef | cochof | cnuevef |sil;
$digito7 = llceroh | llunoh | lldosh | lltresh | llcuatroh | llcincoh |
llseish | llsieteh | llochoh | llnueveh |sil;
$digito8 = llcerof | llunof | lldosf | lltresf | llcuatrof | llcincof |
llseisf | llsietef | llochof |  llnuevef |sil;

$digito9 = oceroh | ounoh | odosh | otresh | ocuatroh | ocincoh | oseish |
osieteh | oochoh | onueveh |sil;
$digito10 = ocerof | ounof | odosf | otresf | ocuatrof | ocincof | oseisf |
osietef | oochof | onuevef |sil;

(
<$digito1>|<$digito2>|<$digito3>|<$digito4>|<$digito5>|<$digito6>|<$digito7
>|<$digito8>|<$digito9>| <$digito10>)
```


SALIDADIALECTOSTEST4

```
---------------------- Sentence Scores ----------------------
====================== HTK Results Analysis ======================
  Date: Wed Mar 20 18:46:05 2002
  Ref : /home/luciano/dialectos/listas/dialectostest4.mlf
  Rec : /home/luciano/dialectos/reconocidos/digitos/MLFsalidadialectostest4
------------------------ File Results --------------------------

                                    .
                                    .
                                    .


a40001c4.rec:   25.00( 12.50)  [H=   2, D=  0, S=  6, I=  1, N=  8]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40001c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40001c4.rec
 LAB: sil         llsieteh llceroh llceroh llsieteh llcincoh llcincoh sil
 REC: sil zsieteh zceroh   zceroh  zceroh  zsieteh  zcincoh  zcincoh  sil
```

```
a40002c1.rec:   88.89( 77.78)  [H=   8, D=  1, S=  0, I=  1, N=  9]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40002c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40002c1.rec
 LAB: sil adosf sil acerof sil          acerof sil atresf sil
 REC:     adosf sil acerof sil acuatrof acerof sil atresf sil
a40002c4.rec:  100.00( 50.00)  [H=   6, D=  0, S=  0, I=  3, N=  6]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40002c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40002c4.rec
 LAB: acerof     acuatrof     acincof        atresf aochof anuevef
 REC: acerof sil acuatrof sil acincof acerof atresf aochof anuevef
a40015c3.rec:   10.00( 10.00)  [H=   1, D=  1, S=  8, I=  0, N= 10]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40015c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40015c3.rec
 LAB: sil acuatrof asietef aseisf acerof   adosf acincof asietef aochof sil
 REC:     zcuatrof zsietef zseisf zcuatrof zdosf zcincof zsietef zochof sil
a40033c3.rec:  100.00( 87.50)  [H=   8, D=  0, S=  0, I=  1, N=  8]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40033c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40033c3.rec
 LAB:     sil csieteh cceroh cochoh ccincoh sil ccuatroh cunoh
 REC: sil sil csieteh cceroh cochoh ccincoh sil ccuatroh cunoh
a40033c4.rec:  100.00( 87.50)  [H=   8, D=  0, S=  0, I=  1, N=  8]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40033c4.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40033c4.rec
 LAB: sil cseish csieteh       cnueveh cnueveh cdosh csieteh sil
 REC: sil cseish csieteh cunoh cnueveh cnueveh cdosh csieteh sil
a40257c1.rec:   33.33( 33.33)  [H=   3, D=  0, S=  6, I=  0, N=  9]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40257c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40257c1.rec
 LAB: sil adosf adosf adosf sil aochof acuatrof acuatrof sil
 REC: sil zdosf zdosf zdosf sil zochof zcuatrof sil     sil
a40258c1.rec:   25.00( 12.50)  [H=   2, D=  0, S=  6, I=  1, N=  8]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40258c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40258c1.rec
 LAB: sil        adosh  adosh anueveh  aunoh  acuatroh  asieteh  sil
 REC: sil lldosh lldosh sil   llnueveh llunoh llcuatroh llsieteh sil
a40258c2.rec:   25.00( 12.50)  [H=   2, D=  0, S=  6, I=  1, N=  8]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40258c2.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40258c2.rec
 LAB: sil        aceroh aceroh adosh sil aochoh aochoh acincoh
 REC: sil ctresh ctresh cceroh cdosh sil cochoh cochoh ccincoh
a40282c1.rec:    0.00(-50.00)  [H=   0, D=  0, S=  6, I=  3, N=  6]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40282c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40282c1.rec
 LAB:                   ztresf  zseisf zdosf   znuevef zcerof  zsietef
 REC: sil otresf oseisf osietef odosf  onuevef ocerof  osietef sil
a40282c2.rec:  100.00(100.00)  [H=   9, D=  0, S=  0, I=  0, N=  9]
a40438c1.rec:  100.00( 20.00)  [H=   5, D=  0, S=  0, I=  4, N=  5]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40438c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40438c1.rec
 LAB: sil llcuatrof           llcerof     llochof     llsietef
 REC: sil llcuatrof lldosf sil llcerof sil llochof sil llsietef
a40438c3.rec:  100.00( 50.00)  [H=   6, D=  0, S=  0, I=  3, N=  6]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40438c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40438c3.rec
 LAB: sil         lltresf     llunof     llseisf llseisf sil
 REC: sil llcerof lltresf sil llunof sil llseisf llseisf sil

a40447c3.rec:  100.00( 71.43)  [H=   7, D=  0, S=  0, I=  2, N=  7]
```

```
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40447c3.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40447c3.rec
 LAB: sil otresf ocerof onuevef        oseisf     ounof osietef
 REC: sil otresf ocerof onuevef otresf oseisf sil ounof osietef
a40462c1.rec:  100.00( 20.00)  [H=   5, D=  0, S=  0, I=  4, N=  5]
Aligned transcription: /home/luciano/etiquetas/digitos/test/a40462c1.lab vs
/home/luciano/dialectos/reconocidos/digitos/a40462c1.rec
 LAB: sil llcuatrof             llunof        llnuevef     llcerof
 REC: sil llcuatrof llcerof sil llunof lldosf llnuevef sil llcerof
------------------------ Overall Results --------------------------
SENT: %Correct=0.72 [H=1, S=137, N=138]
WORD: %Corr=34.89, Acc=18.92 [H=389, D=54, S=672, I=178, N=1115]
========================================================================
```