UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

TRANSCRIPCION AUTOMATICA DE VOZ ESPONTANEA PARA RECUPERACIÓN DE INFORMACION

Sergio Lucas Bermejo

Septiembre 2008

PROYECTO FIN DE CARRERA

Título: *Transcripción automática de voz espontánea para recuperación de información*

Autor: D. Sergio Lucas Bermejo

Tutor: D. Doroteo Torre Toledano

Tribunal:

Presidente: Joaquín González Rodríguez

Vocal: Pablo Castells Azpilicueta

Vocal secretario: Doroteo Torre Toledano

Fecha de lectura:

Calificación:

TRANSCRIPCION AUTOMATICA DE VOZ ESPONTANEA PARA RECUPERACIÓN DE INFORMACION

AUTOR: Sergio Lucas Bermejo
TUTOR: Doroteo Torre Toledano

Área de Tratamiento de Voz y Señales
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre de 2008

Resumen:

En este proyecto se estudian los sistemas de reconocimiento de habla inglesa, haciendo uso de las herramientas proporcionadas por HTK y del transcriptor fonético de palabras para inglés americano implementado en el desarrollo del proyecto. Para el entrenamiento de los modelos acústicos y la evaluación de los sistemas de reconocimiento, se ha utilizado la base de datos de habla telefónica espontánea *SpeechDat-English*. En cuanto al modelado acústico, se han construido modelos dependientes e independientes de contexto mediante modelos Ocultos de Markov (HMMs).

En la parte experimental se han desarrollado pruebas de reconocimiento fonético y de reconocimiento de palabras, haciendo uso tanto de los modelos dependientes de contexto como de los independientes de contexto, con el fin de evaluar los distintos sistemas y extraer las conclusiones pertinentes.

Palabras clave:
Sistema de reconocimiento de habla, Reconocimiento fonético, Reconocimiento de palabras, transcriptor fonético de palabras, Modelos Ocultos de Markov, habla telefónica espontánea, *SpeechDat*, HTK.

Abstract:

This project will study the systems of English recognition, using the tools provided by HTK and the phonetic transcriptor of words to American English implemented in the development of the project. For the training of acoustic models and the evaluation of recognition systems, the database of telephone spontaneous speech has been used, SpeechDat-English. For the acoustic modeling, context-dependent and context-independent models have been built by Hidden Markov models (HMMs).

In the experimental tests, phonetic recognition and word recognition have been developed making use of both, the context-dependent and the context-independent models, in order to assess the various systems and draw the appropriate conclusions.

Keywords

Speech recognition system, phonetic recognition, word recognition, phonetic transcriber of words, Hidden Markov Models, telephone spontaneous speech, SpeechDat, HTK

Agradecimientos

Quiero agradecer en primer lugar a mi tutor, Doroteo Torre Toledano, por la oportunidad que me ha brindado de realizar el Proyecto de Fin de Carrera en el grupo de investigación ATVS, así como por su dedicación y apoyo en el transcurso del mismo.

Además, el desarrollo de este proyecto no hubiese sido posible sin la cooperación de Joaquín González Rodríguez y del resto de miembros del ATVS. Agradezco también el apoyo de Daniel Ramos, Alejandro Abejón, Javier González, Javier Franco, Ismael Mateos, Daniel Hernández, Ignacio López, Víctor González, que de alguna manera u otra, me han ayudado y animado durante estos meses.

También, quisiera agradecer a José María Martínez y a Jesús Bescós el tiempo que han empleado en escucharme y orientarme a lo largo de estos años de estudios en la Escuela Politécnica Superior (EPS) de la Universidad Autónoma de Madrid.

En mis 5 años de estudios en la EPS, he conocido a personas muy especiales, con las que he pasado muy buenos momentos, bueno… y algún que otro momento de agobio también. Quiero agradecer esos momentos a Bárbara Valenciano, Sonsoles Herrero, Verónica Peña, Elena Ortiz, Cristina Monsalve, Marcos Martínez, Javier Castillo, Alberto Harriero y Pedro Tomé.

Una beca Erasmus en Berlín de 10 meses, me dio la oportunidad de conocer a personas que han llegado a ser muy importantes en mi vida y que son responsables de que no olvide jamás esos meses. Por ello, quiero mencionar a María, Lucia, Tanja, Laia, Lana, Andrea, Michi, Adrián, Miguel, Javier, Enzo, Payés, Manuel, Arda y George.

Como no agradecer a mis amigos más viejos: Cristina Rodrigo, Itziar Esteban-Infantes, Carmen de Paz, Aurora Macías, Paula Martínez, Inés Benezet, Ana Acero, Gema de Lama, Patricia Perulero, Isa González, María García, Juan Ignacio Merino, Fernando Sánchez, Pedro Antonio Gómez, Mariano García, Pedro Sánchez, Luis Perezagua y Carlos Valcárcel, con los que he crecido y crezco, todos los momentos que hemos pasado y lo que han aportado todos estos años. Y en especial a Carlos, por obligarme a madrugar esos días de Agosto y hacer de su salón una biblioteca, posibilitando que este proyecto viese la luz en su fecha.

Por supuesto, quiero dedicar este proyecto a mis padres, Nicolás y Mara, a mi hermana, Marien, y a mi abuela, Herminia, por todo el apoyo, cariño y confianza que me han dado a lo largo de toda mi vida, y que sin duda, han sido personas claves en la culminación de mis estudios (y no sólo por subvencionarme las matrículas de la carrera ☺).

Y por último, deseo a todas esas personas que tengan el valor, las ganas y el tiempo para pasar de esta hoja, que les sea útil la lectura de este proyecto.

Sergio Lucas Bermejo,
Septiembre de 2008.

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica superior de la Universidad Autónoma de Madrid. El proyecto ha sido financiado parcialmente por el Ministerio de Defensa y el Ministerio de Educación y Ciencia con el proyecto TEC2006-13170-C02-01.

INDICE DE CONTENIDOS

1. Introducción ... 1
1.1. Motivación ... 1
1.2. Objetivos ... 2
1.3. Organización de la memoria ... 2
2. Estado del arte ... 3
2.1. Introducción. ... 3
2.1.1. Procesamiento y Caracterización de la señal. ... 4
2.1.2. Extracción de parámetros: MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs). ... 5
2.1.3. Clasificación de Patrones. ... 7
2.1.4. Decisión. ... 7
2.2. Sonido y procesamiento humano del habla. ... 8
2.2.1. Introducción. ... 8
2.2.2. Sonido. ... 8
2.2.3. Producción del habla. ... 9
2.2.4. Percepción del habla. ... 11
2.2.4.1. Análisis espectral. ... 13
2.3. Fonética y fonología. ... 14
2.3.1. Introducción. ... 14
2.3.2. Fono y fonema. ... 15
2.3.2.1. Características de un fonema. ... 16
2.4. Creación de modelos fonéticos. ... 16
2.4.1. Hidden Markov Models (HMMs). ... 19
2.4.1.1. Introducción. ... 19
2.4.1.2. Elementos de un HMM. ... 22
2.4.1.3. Problemas a resolver para la utilización de un HMM. ... 23
2.5. Modelado de idioma. ... 30
2.5.1. Modelos de lenguaje de n-gramas. ... 31
2.6. Reconocimiento basado en N-Best y lattices. ... 32
3. Diseño y desarrollo. ... 35
3.1. Medios disponibles. ... 35
3.1.1. Bases de datos. ... 35
3.1.2. SpeechDat. ... 35
3.1.2.1. Preparación de los datos. ... 37
3.1.3. Software. ... 38
3.1.4. Hardware. ... 39
3.2. Diseño. ... 39
3.2.1. Transcriptor fonético de palabras para inglés americano. ... 39
3.2.2. Parametrización de las señales de audio. ... 42
3.2.3. Creación de los modelos acústicos independientes del contexto. ... 43
3.2.4. Creación de los modelos acústicos dependientes del contexto - Tri-fonemas. ... 45
3.2.5. Construcción de la gramática. ... 46
3.2.6. Sistema de reconocimiento fonético. ... 47
3.2.6.1. Obtención de las transcripciones fonéticas. ... 47
3.2.6.2. Entrenamiento de los modelos acústicos fonéticos. ... 47
3.2.6.3. Construcción de la gramática basada en fonemas. ... 47
3.2.7. Sistema de reconocimiento de palabras basado en modelos fonéticos. ... 48
3.2.7.1. Gramática basada en palabras. ... 48
3.2.8. Sistema de reconocimiento de palabras basado en modelos de tri-fonemas. ... 48
3.2.8.1. Obtención de las transcripciones a nivel de tri-fonemas. ... 48
3.2.8.2. Entrenamiento de los modelos acústicos basados en tri-fonemas. ... 48
3.2.9. Evaluación de los modelos acústicos. ... 50
4. Integración, pruebas y resultados ... 53

4.1. Reconocimiento fonético. 53
4.2. Reconocimiento a nivel de palabras. 54
4.2.1. Reconocimiento basado en modelos fonéticos. 54
4.2.2. Reconocimiento basado en modelos de tri-fonemas. 55
4.2.3. Reconocimiento con N-best. 57
4.3. Otros experimentos. 61
5. Conclusiones y trabajo futuro 63
5.1. Conclusiones 63
5.2. Trabajo futuro 64
Referencias 65
Glosario LXIX

INDICE DE FIGURAS

FIGURA 1: ARQUITECTURA BÁSICA DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA, DERIVADO DE LAS FIGURAS ENCONTRADAS EN [JUNQA Y HATON, 1996; P.84] Y [HUANG ET AL., 2001; P.5]. 3

FIGURA 2- SEÑAL DE VOZ Y SU CORRESPONDIENTE ESPECTROGRAMA DE ANCHO DE BANDA [HUANG ET AL., 2001; P. 277]. LAS ÁREAS MÁS OSCURAS SIGNIFICAN MAYOR ENERGÍA PA RA ESE TIEMPO Y ESA CRECENCIA. 4

FIGURA 3- DIAGRAMA DE BLOQUES DEL PROCESO DE CÁLCULO DE LOS MEL-FREQUENCY CEPSTRAL COEFFICIENTS. 5

FIGURA 4- ESQUEMATIZACIÓN DE LOS DELTA-MEL-FREQUENCY CEPSTRAL COEFFICIENTS. 6

FIGURA 5- LA APLICACIÓN DE ENERGÍA PROVOCA ALTERNATIVAMENTE COMPRESIÓN Y REFRACCIÓN DE MOLÉCULAS DE AIRE, QUE DESCRIBE UNA ONDA SINUSOIDAL. [HUANG ET AL., 2001; P. 21]. LAS ÁREAS MÁS OSCURAS SIGNIFICAN MAYOR CONCENTRACIÓN DE MOLÉCULAS DE AIRE. 8

FIGURA 6- EL NIVEL SPL EN DB DEL UMBRAL ABSOLUTO DE AUDICIÓN EN FUNCIÓN DE LA FRECRECENCIA [HUANG ET AL., 2001; P. 23]. LOS SONIDOS POR DEBAJO DE ESTE NIVEL SON INAUDIBLES. NOTE QUE ANTES DE 100HZ Y DESPUÉS DE 10KHZ ESTE LÍMITE SE ALCANZA RÁPIDAMENTE. 9

FIGURA 7- SECCIÓN SAGITAL DE LA CAVIDAD ORAL [HUANG ET AL., 2001; P. 24]. 10

FIGURA 8- CICLO DE LAS CUERDAS VOCALES EN LA LARINGE [HUANG ET AL., 2001; P. 26]. (A) CUERDAS VOCALES CERRADAS (INCREMENTO DE LA PRESIÓN SUBGLOTAL). (B) APERTURA DE LAS CUERDAS VOCALES DEBIDO A LA ALTA PRESIÓN SUBGLOTAL ALCANZADA (DISMINUCIÓN DE LA PRESIÓN SUBGLOTAL). (C) LA FUERZA DE LOS MÚSCULOS CIERRAN LAS CUERDAS VOCALES COMO CONSECUENCIA DE LA DISMINUCIÓN DE LA PRESIÓN (COMIENZO DEL PRÓXIMO CICLO). 11

FIGURA 9- FLUJO DE AIRE DURANTE EL CICLO DE LA LARINGE [HUANG ET AL., 2001; P. 27]. 11

FIGURA 10- ESTRUCTURA DEL SISTEMA PERIFÉRICO AUDITIVO [HUANG ET AL., 2001; P. 29]. 12

FIGURA 11- FRECUENCIA CENTRAL DE LOS 24 FILTROS BARK [HUANG ET AL., 2001; P. 33]. 14

FIGURA 12- FUNCIÓN DE ADAPTACIÓN DTW. 17

FIGURA 13- VQ BIDIMENSIONAL. 18

FIGURA 14- ESQUEMA DE UNA CADENA DE MARKOV PARA LA COTIZACIÓN EN EL DOW JONES [HUANG ET AL., 2001; P. 381]. 20

FIGURA 15- ESQUEMA DE UN HMM PARA LA COTIZACIÓN EN EL DOW JONES [HUANG ET AL., 2001; P. 381]. 21

FIGURA 16- EL MODELO DE GENERACIÓN DE MARKOV[THE HTK BOOK, 2005]. 22

FIGURA 17- RELACIÓN DE α_{t-1} Y α_t Y DE β_t Y β_{t+1} EN EL ALGORITMO FORWARD-BACKWARD [HUANG ET AL., 2001; P. 390]. 26

FIGURA 18- ILUSTRACIÓN DE LAS OPERACIONES NECESARIAS PARA EL CÁLCULO DE $\xi_t(i,j)$ [HUANG ET AL., 2001; P. 391]. 28

FIGURA 19- REPRESENTACIÓN GRÁFICA DEL LATTICE DE PALABRAS, PARA LA FRASE DEL ÁMBITO DE NEGOCIOS. 33

FIGURA 20- DIAGRAMA DE BLOQUES QUE REPRESENTA EL FUNCIONAMIENTO DEL TRANSCRIPTOR FONÉTICO DE PALABRAS. 42

FIGURA 21- TOPOLOGÍA DEL MODELO ACÚSTICO DE LOS FONEMAS. 43

FIGURA 22- FORMATO HTK DEL MODELO ACÚSTICO DE LOS FONEMAS. 44

FIGURA 23- TOPOLOGÍA DEL MODELO ACÚSTICO PARA LA PAUSA CORTA. 44

FIGURA 24- TOPOLOGÍA DEL MODELO DE SILENCIO [HTK BOOK, P.33]. 45

FIGURA 25- MATRICES DE TRANSICIÓN COMPARTIDAS [HTK BOOK, P.36]. 46

FIGURA 26- ESPECTROGRAMA PARA EL FONEMA /IY/ CON DOS FONEMAS VECINOS A LA IZQUIERDA DISTINTOS: /R/ Y /W/ [HUANG ET AL., 2001; P. 432]. 49

FIGURA 27- FORMATO ESTANDARIZADO POR NIST PARA LA REPRESENTACIÓN DE RESULTADOS DE RECONOCIMIENTO DE HABLA. 51

FIGURA 28- DIAGRAMA DE BLOQUES DEL PROCESO DE ENTRENAMIENTO Y EVALUACIÓN DE LOS MODELOS ACÚSTICOS. 52

FIGURA 29- RESULTADOS OBTENIDOS EN RECONOCIMIENTO FONÉTICO CON LOS MODELOS FONÉTICOS DE 39 GAUSSIANAS OBTENIDOS. 53

FIGURA 30- ALGUNAS ENTRADAS DEL DICCIONARIO FONÉTICO. 54

FIGURA 31- RESULTADOS OBTENIDOS EN RECONOCIMIENTO A NIVEL DE PALABRAS CON LOS MODELOS FONÉTICOS DE 39 GAUSSIANAS OBTENIDOS. 55

FIGURA 32- ALGUNAS ENTRADAS DEL DICCIONARIO A NIVEL DE TRI-FONÉTICO. 56

FIGURA 33- RESULTADOS OBTENIDOS EN RECONOCIMIENTO A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN TRI-FONEMAS DE 39 GAUSSIANAS OBTENIDOS. 56

FIGURA 34- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 100-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN FONEMAS DE 39 GAUSSIANAS. 57

FIGURA 35- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 75-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN FONEMAS DE 39 GAUSSIANAS. 57

FIGURA 36- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 50-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN FONEMAS DE 39 GAUSSIANAS. 58

FIGURA 37- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 25-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN FONEMAS DE 39 GAUSSIANAS. 58

FIGURA 38- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 100-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN TRI-FONEMAS DE 39 GAUSSIANAS. 58

FIGURA 39- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 75-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN TRI-FONEMAS DE 39 GAUSSIANAS. 59

FIGURA 40- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 50-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN TRI-FONEMAS DE 39 GAUSSIANAS. 59

FIGURA 41- RESULTADOS OBTENIDOS EN RECONOCIMIENTO 25-BEST A NIVEL DE PALABRAS CON LOS MODELOS BASADOS EN TRI-FONEMAS DE 39 GAUSSIANAS. 59

INDICE DE TABLAS

TABLA 1- EJEMPLO DE UNA LISTA 10-BEST PARA UNA FRASE DEL ÁMBITO DE LOS NEGOCIOS [HUANG ET AL., 2001; P. 665] 32

TABLA 2- DISTRIBUCIÓN DE LAS EDADES DE LOS LOCUTORES. 35

TABLA 3- FORMATO DEL ARCHIVO DE DATOS ASOCIADO A UN ARCHIVO DE AUDIO [SPEECHDAT ENGLISH]. 36

TABLA 4- TIPOS DE RUIDOS MODELADOS Y SU FRECUENCIA. 38

TABLA 5- CONJUNTO DE FONEMAS ELEGIDOS. 39

TABLA 6- REPRESENTACIÓN DE LOS CARACTERES IPA CON EL CONJUNTO DE FONEMAS DEL DISEÑO. 40

TABLA 7- CARACTERÍSTICAS DE LOS FONEMAS QUE REPRESENTAN VOCALES DEL HABLA INGLESA. 50

TABLA 8- CONSONANTES DEL HABLA INGLESA SEGÚN EL LUGAR DE ARTICULACIÓN (COLUMNAS) Y LA CLASE DE ARTICULADOR (FILAS). 50

TABLA 9- CLASIFICACIÓN DE LAS CONSONANTES DEL HABLA INGLESA. 50

1. Introducción

1.1. Motivación

El Reconocimiento Automático del habla (RAH) tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras electrónicas y el procesamiento de información grabada de forma automática. El problema que se plantea en un sistema de RAH es hacer cooperar un conjunto de informaciones provenientes de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido [F. Casacubierta, E. Vidal, 1987].

Las aplicaciones de reconocimiento de habla incluyen, entre otras, llamada por voz, domótica, búsqueda, entrada de datos (como introducción del número de la tarjeta de crédito), preparación de documentos estructurados, *Word-Spotting*, transcripción automática y en las cabinas de pilotos de los aviones.

Los sistemas comerciales han estado disponibles desde 1990 y el campo en el que está siendo más utilizado es en el de aplicaciones telefónicas: agencias de viajes, atención al cliente, información, etc. La mejoría de estos sistemas de reconocimiento del habla han ido aumentando y su eficiencia cada vez es mayor. En concreto, los sistemas de transcripción automática de voz están cobrando gran importancia, debido a las múltiples aplicaciones como *Word-Spotting* o la búsqueda de videos y archivos de audio por contenido, ya que a la hora de hacer una búsqueda en contenidos de audio y audiovisuales es muy útil partir de una transcripción previa para hacer posteriormente búsquedas sobre dichas transcripciones

Como curiosidad, comentar que los investigadores del grupo de reconocimiento de voz de Apple solían llevar una camiseta en la que se podía leer *'I helped Apple wreck a nice Beach'* (ayudé a Apple a estropear una bonita playa), cuya pronunciación es idéntica a *'I helped Apple recognize speech'* (ayudé a Apple a reconocer habla). Esta broma ilustra la dificultad de desambiguar cadenas fonéticas.

1.2. Objetivos

El objetivo de este proyecto es construir un sistema de transcripción automática de voz espontánea para recuperación de información de audio y audiovisual.

La implementación del sistema está basada en la técnica de modelado estadístico de los modelos ocultos de Markov o HMMs *(Hidden Markov Models)*. Para el entrenamiento de dichos modelos y para la evaluación del sistema, se hace uso de la base de datos *SpeechDat-English* de habla telefónica.

Finalmente, con los modelos acústicos obtenidos, se realizaran distintas pruebas con el fin de evaluar el sistema implementado y extraer conclusiones de las mismas.

1.3. Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Estado del arte** en reconocimiento fonético, introduciendo conceptos sobre el sonido, la fonética y un planteamiento general del reconocimiento del habla, para posteriormente, hacer un estudio de las distintas técnicas existentes, profundizando especialmente en la técnica usada en el desarrollo del proyecto, los *Hidden Markov Models* (HMMs).
- **Diseño y desarrollo** del proyecto, describiendo los medios disponibles desde los cuales se partió (Bases de datos, *Software* y *Hardware*) y los distintos pasos seguidos para la implementación del sistema.
- **Integración, pruebas y resultados** obtenidos de las mismas, para la evaluación, análisis y extracción de conclusiones del sistema implementado.
- **Conclusiones y trabajo futuro** que resulte interesante a la vista de los resultados obtenidos para mejorar el sistema implementado.
- **Referencias** de las distintas fuentes de información utilizadas para el desarrollo de esta memoria**.**
- **Glosario.**

2. Estado del arte

En esta Sección se va a realizar un análisis y clasificación de los sonidos, así como el estudio de las técnicas de modelado acústico y de idioma para su aplicación en los reconocedores de voz.

2.1. Introducción.

El modelado acústico incluye el conocimiento acerca de la acústica, fonética, variabilidad del entorno, sexo, diferencias entre los distintos dialectos de diferentes locutores, etc. Mientras que el modelado de idioma está referido al conocimiento semántico, sintáctico y pragmático, es decir, al conocimiento de las posibles palabras, su probabilidad de ocurrencia en una secuencia determinada [N. Morales, 2007, pp. 5-7].

La estructura de los sistemas de reconocimiento del habla (RAH) se basa, generalmente, en tres etapas (véase la *Figura 1*): procesamiento y caracterización de la señal, clasificación de patrones y decisión. Estás tres etapas son brevemente comentadas en las siguientes Secciones.

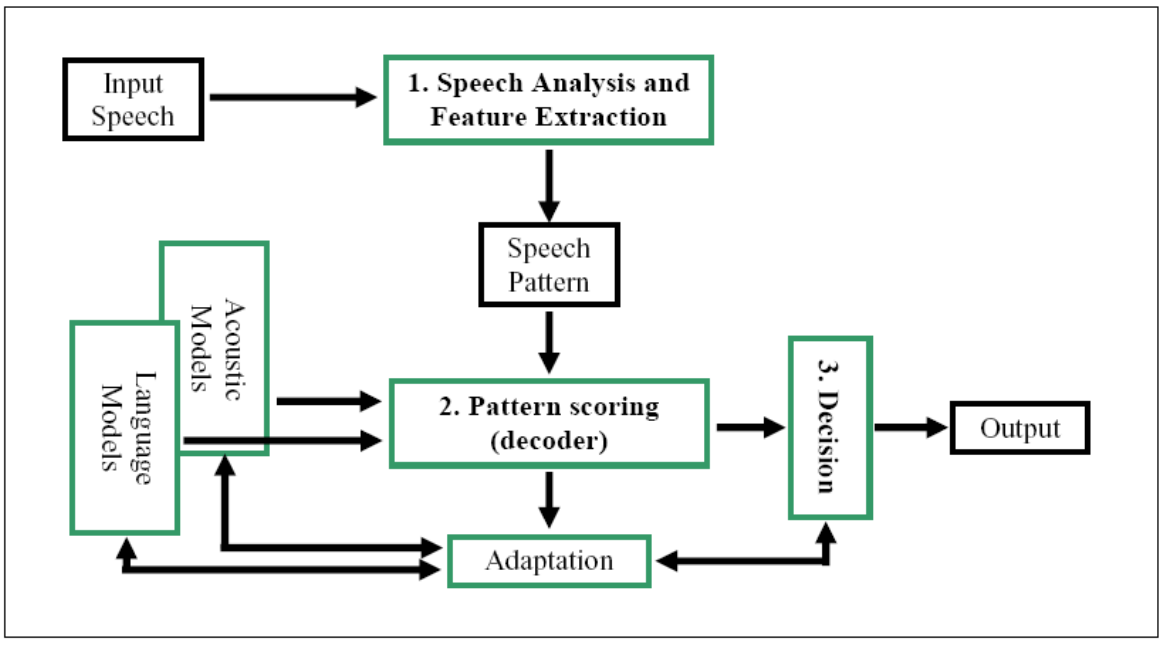

Figura 1: Arquitectura básica de un sistema de Reconocimiento Automático del Habla, derivado de las figuras encontradas en [Junqua y Haton, 1996; p.84] y [Huang *et al.*, 2001; p.5].

2.1.1. Procesamiento y Caracterización de la señal.

Los pasos que se realizan sobre la señal de voz, de principio a fin, son los siguientes:

- La señal de voz es muestreada periódicamente con una tasa constante, que excede la tasa de muestreo de Nyquist (Teorema de Nyquist [Nyquist, 1928] [Shannon, 1949]), evitando así el solapamiento.

- En la práctica, es muy común aplicar una técnica de mejora llamada pre-énfasis, que consiste en aumentar las altas frecuencias de la señal de voz con el fin de compensar el desplazamiento espectral producido en la señal acústica como resultado de la radiación de salida desde los labios. Esta técnica también tiene lugar el oído humano [Moore, 2003; pp. 55-59] para suavizar el espectro y generalmente se tiene en cuenta para mejorar la precisión del RAH.

- La señal temporal es divida en una secuencia de ventanas de corta duración, cada una de las cuales contiene una parte de la señal cuasi-estacionaria (véase la *Figura 2*) (correspondiente a configuraciones cuasi-estáticas del tracto vocal). Con el objetivo de evitar los efectos de borde, se utilizan funciones de enventanado que se aproximen a cero cerca de los bordes (Hamming, Hanning,…).También es usual permitir el solapamiento entre ventanas consecutivas con el fin de obtener una mayor resolución temporal (siendo el valor típico de la anchura de las ventanas del orden de 25 ms y el desplazamiento de las ventanas del orden de 10 ms).

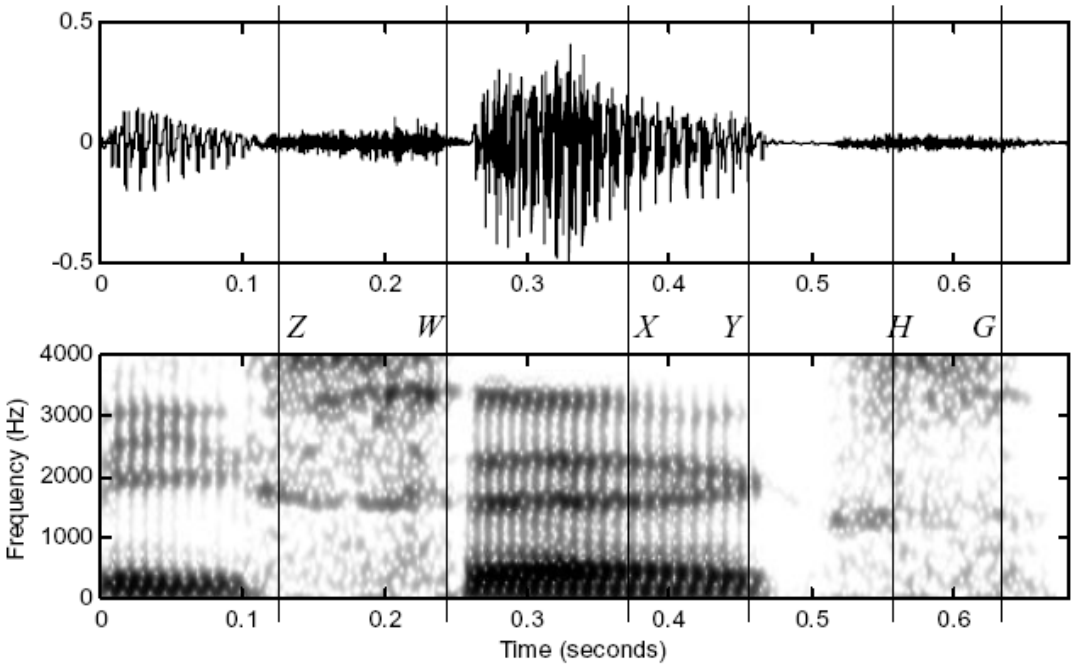

Figura 2- Señal de voz y su correspondiente espectrograma de ancho de banda [Huang *et al.*, 2001; p. 277]. Las áreas más oscuras significan mayor energía para ese tiempo y esa frecuencia.

2.1.2. Extracción de parámetros: MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs).

Existen numerosos experimentos sobre la audición humana que demuestran que el oído interno (cóclea) actúa como un analizador espectral (véase la *Sección 2.2.4.1*). Por otra parte, análisis sobre la producción humana del habla muestran que los locutores tienden a controlar mucho más el contenido espectral de su salida que los detalles de las formas de onda del habla de salida. Por lo que obviamente la extracción de los parámetros se lleva a cabo en el dominio de la frecuencia.

El método de parametrización más común en el reconocimiento de habla es del tipo ***Mel-frequency cepstral coefficients (MFCCs)***. En la *Figura 3*, se muestra el esquema de extracción de los MFCCs.

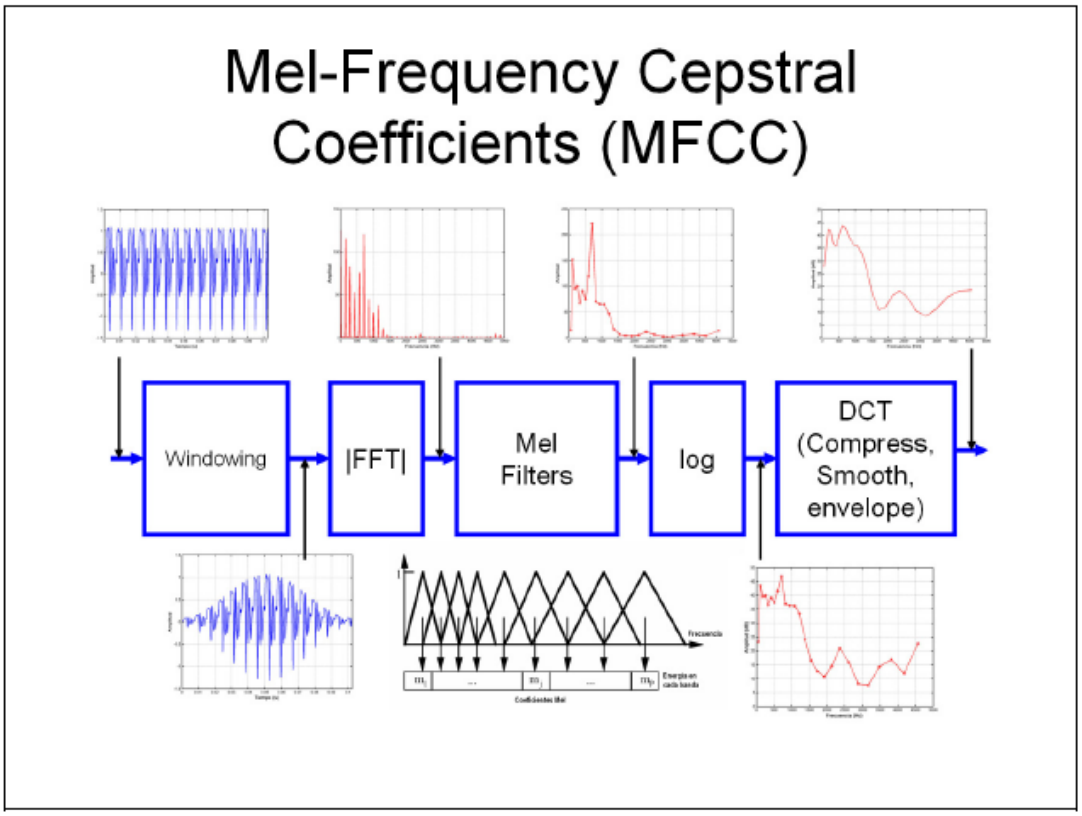

Figura 3- Diagrama de bloques del proceso de cálculo de los Mel-Frequency Cepstral Coefficients.

El proceso de extracción de los MFCCs se puede resumir en los siguientes pasos:

1. **Enventanar** la señal en segmentos de 25 ms, con un desplazamiento de 10 ms entre ventanas, como ya se ha comentado.
2. **Pasar al dominio de la frecuencia** cada segmento por medio de la *Fast Fourier Transform* (FFT).
3. **Aplicar un banco de filtros** de diferentes frecuencias y amplitudes, dando mayor resolución a las bajas frecuencias. Con ello simulamos el oído interno (cóclea) del sistema auditivo humano.
4. **Calcular la en energía en promedio** de la salida de cada uno de los filtros. Los valores obtenidos se pueden ver como una nueva señal de tiempo discreto. Así

por ejemplo, usando un banco de 40 filtros, obtendríamos un vector de 40 coeficientes.

5. **Transformar a través de la *Discrete Cosine Transform* (DCT)** la señal en tiempo discreto obtenida. Así se obtienen una serie de parámetros (habitualmente de 13 a 20) aproximadamente incorrelados entre sí. Estos parámetros son los MFCCs.

Los MFCCs representan la envolvente espectral de la señal de voz, obteniendo así de ellos importantes características del habla. En concreto, el primer coeficiente, C0, indica la energía de la señal y se usa o no dependiendo de la aplicación. Y el segundo coeficiente, C1, tiene una razonable interpretación como indicador del balance global de energía entre bajas y altas frecuencias [D. O'Shaughnessy, 2008, p. 2973].

Para obtener más información relevante, como la coarticulación de los fonemas, se suelen usar también las velocidades (Delta-MFCC) y/o las aceleraciones (Delta-Delta-MFCC) que representan la evolución temporal de los fonemas en su transición a otros fonemas.

Los coeficientes Delta representan la variación de los MFCCs alrededor del instante de tiempo considerado. Por ello, son denominados coeficientes de velocidad o de primera derivada (*Figura 4*). Los coeficientes Delta-Delta representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración.

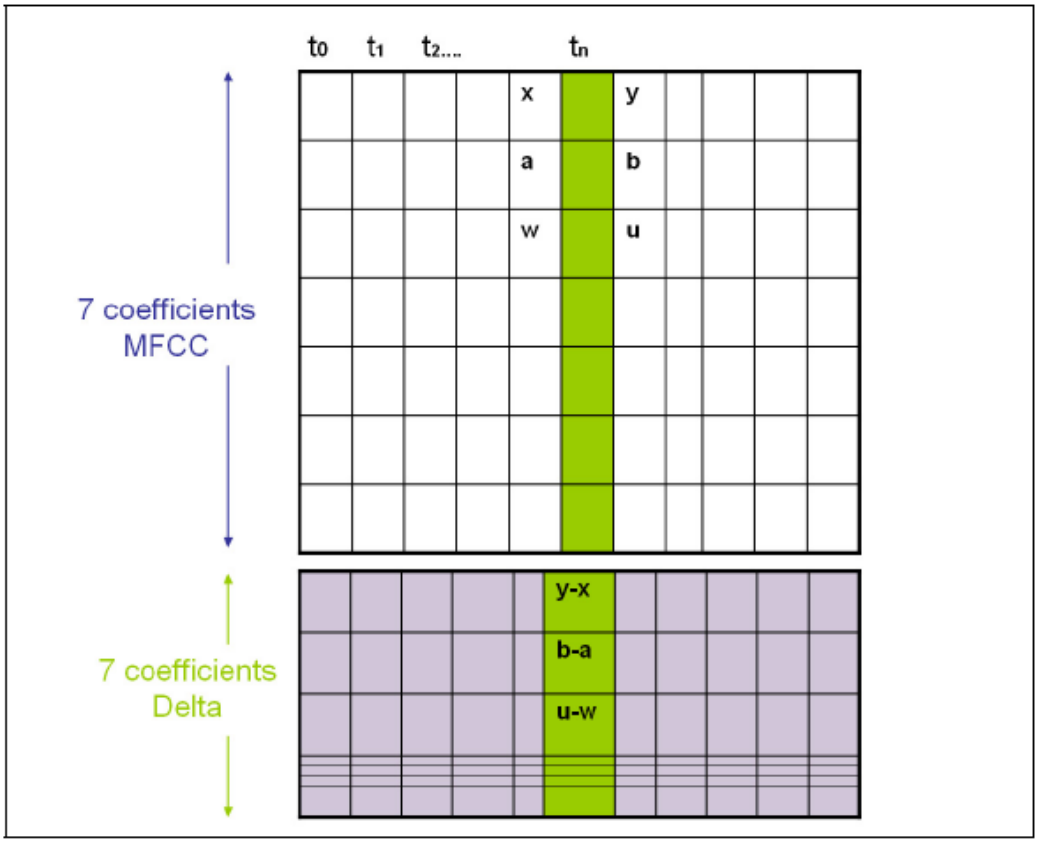

Figura 4- Esquematización de los Delta-Mel-frequency Cepstral Coefficients.

Sin embargo, los MFCCs son difíciles de relacionar con cualquier aspecto cerrado de la producción o percepción del habla. Los detalles espectrales que contienen, permiten la discriminación entre sonidos similares, pero su carencia de interpretación los hace altamente vulnerables a condiciones no lineales tales como el ruido o acentos. En particular, los MFCCs dan igual peso a las altas y bajas amplitudes en el espectro logarítmico, cuando es bien conocido que la alta energía domina en la percepción.

2.1.3. Clasificación de Patrones.

En este módulo, se comparan las distintas ventanas o partes de la señal con los modelos acústicos del sistema y se obtiene la secuencia (o secuencias) de estados más probable para un sonido. Existen diferentes métodos de clasificación, sin embargo los modelos acústicos más usados en los actuales sistemas son los Hidden Markov Models (HMMs), que representan la unidad mínima de habla (fonemas, sílabas, palabras, etc) como una sucesión de estados, cada uno de los cuales está caracterizado por una probabilidad de emisión (modelada como una función de densidad de probabilidad, fdp) y una probabilidad de transición a otros estados. Los HMMs se discuten en la *Sección 2.4.1*.

Debido a que el entendimiento humano del habla no se basa solamente en la información acústica (contexto del habla, características del locutor, etc), la clasificación de patrones del sistema no debería estar limitada por la información acústica. Por ello, para dotar al sistema de mayor robustez, se introduce información a diferentes niveles, como por ejemplo, información sobre el idioma (conocimiento semántico, sintáctico y pragmático) mediante los modelos de lenguaje (LMs). Los LMs se discuten en la Sección *2.5*.

Adicionalmente, como se puede observar en la *Figura 1*, el sistema incluye un módulo de adaptación de los modelos acústicos y del modelo de lenguaje. Dicha adaptación, se puede llevar a cabo durante el reconocimiento. Por ejemplo, los modelos acústicos pueden ser adaptados para ajustarse a las características particulares de un determinado locutor o para compensar la distorsión. Y los LMs pueden ser simplificados significativamente si se identifica el tema del mensaje.

2.1.4. Decisión.

La secuencia (o secuencias) de estados más probable es pasada al módulo de decisión, en la que se llevan a cabo diversas acciones para decidir que transcripción es la más acertada, según la secuencia de observaciones acústicas de entrada y los modelos acústicos y de idioma. La decisión puede ser tan simple como aceptar la transcripción textual más probable. También es posible el post-procesamiento, obteniendo mayor precisión o experiencia de usuario. Otra opción, sería el procesamiento de varios candidatos (*lattices* o listas *N-best*, [Richardson *et al.*, 1995] [Nguyen *et al.*, 1994]), que será estudiado con más detalle en la *Sección 2.6*.

2.2. Sonido y procesamiento humano del habla.

2.2.1. Introducción.

Antes de explicar el estado del arte de las distintas técnicas en reconocimiento fonético, vamos a describir y clasificar los distintos patrones de sonidos de un idioma hablado, así como una breve explicación del sistema humano de producción y percepción del habla.

Las señales de habla están compuestas de patrones de sonidos analógicos, que sirven como base para una representación discreta y simbólica de un idioma hablado – fonemas, sílabas y palabras. La producción e interpretación de estos sonidos se rigen por la sintaxis y semántica del idioma hablado [Huang *et al.*, 2001, pp. 21-36].

2.2.2. Sonido.

Un sonido es una onda de presión longitudinal formada por compresiones y expansiones del aire, en dirección paralela a la aplicación de energía. Las compresiones son zonas donde las moléculas de aire han sido forzadas por la aplicación de energía, dando lugar a una mayor concentración de las mismas. Y las expansiones son zonas donde la concentración de moléculas de aire es menor.

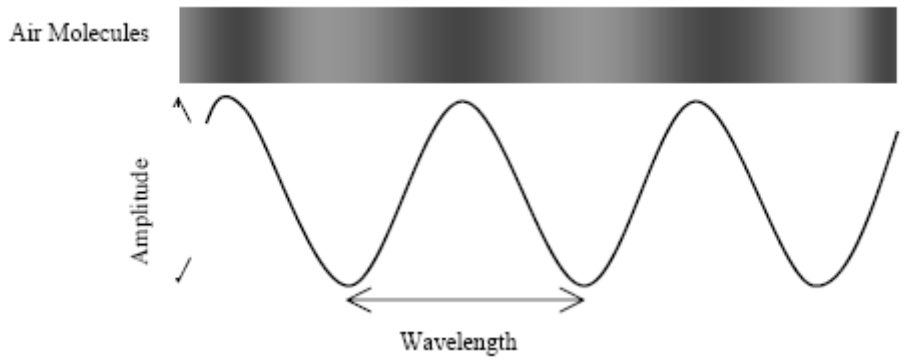

Figura 5- La aplicación de energía provoca alternativamente compresión y expansión de moléculas de aire, que describe una onda sinusoidal. [Huang *et al.*, 2001; p. 21]. Las áreas más oscuras significan mayor concentración de moléculas de aire.

Hay dos parámetros importantes para describir una onda sinusoidal, amplitud y longitud de onda (la frecuencia es usada también para la descripción de una onda sinusoidal).

La cantidad de trabajo realizado para generar la energía que produce la compresión de moléculas de aire, se refleja en la cantidad de desplazamiento de las moléculas de aire desde su posición de reposo. Este grado de desplazamiento es medido como la amplitud de un sonido, como se muestra en la *Figura 5*. Debido al amplio rango de variación, es conveniente medir la amplitud del sonido en decibelios (dB) sobre una escala logarítmica.

El nivel de presión de un sonido (SPL, Sound Pressure Level) es una medida de la presión del sonido, $\boldsymbol{P}$ en dB:

$$SLP(dB) = 20\log_{10}\left(\frac{P}{P_0}\right)$$

donde la referencia 0 dB corresponde al umbral de audición del oído humano, el cual es P_0= 0.0002 μbar para un tono de 1 Khz, de manera que los sonidos con un nivel SPL menor a 0 dB son inaudibles para el oído humano. Por otro lado, los sonidos más altos que el oído humano puede tolerar son en torno a los 120 dB, denominado como el umbral de dolor de audición.

El umbral absoluto de audición es la máxima cantidad de energía de un tono puro que no puede ser detectado por un oyente en un entrono sin ruido. Dicho umbral es función de la frecuencia y puede ser aproximado por:

$$T_q(f) = 3.64(f/100)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \qquad (dB\ SPL)$$

y es representada en la *Figura 6*.

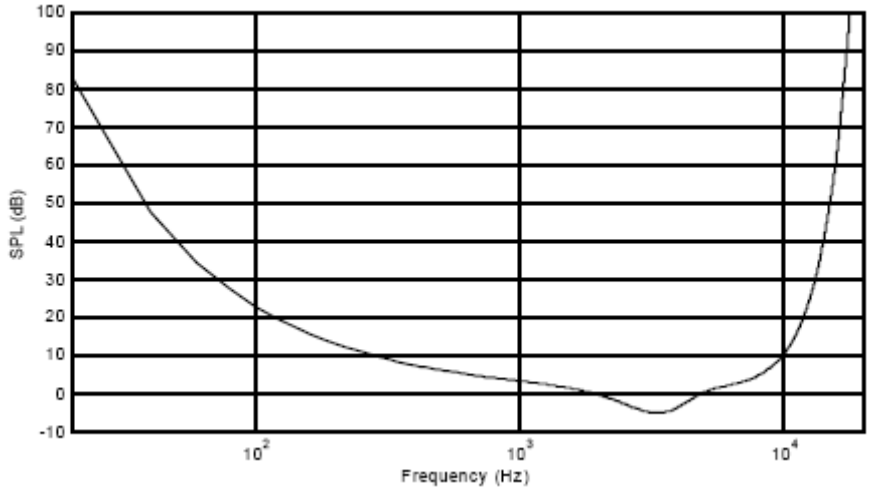

Figura 6- El nivel SPL en dB del umbral absoluto de audición en función de la frecrecencia [Huang *et al.*, 2001; p. 23]. Los sonidos por debajo de este nivel son inaudibles. Note que antes de 100Hz y después de 10kHz este límite se alcanza rápidamente.

2.2.3. Producción del habla.

El habla es producida por ondas de presión de aire provenientes de la boca y la ventana de la nariz del locutor.

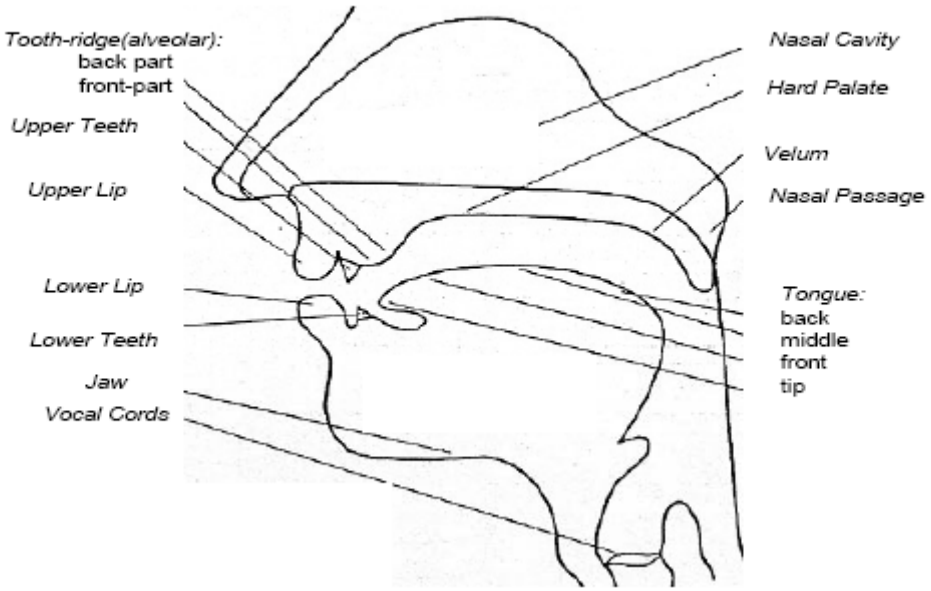

Figura 7- Sección sagital de la cavidad oral [Huang *et al.*, 2001; p. 24].

La cavidad laríngea es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo (o no, como veremos), en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas.
Las cavidades supraglóticas están constituidas por la faringe (garganta), la cavidad nasal y la cavidad bucal. La faringe y la cavidad bucal son referidas como el tracto vocal y la cavidad nasal como el tracto nasal. Como se ilustra en la *Figura 7*, el aparato de producción de habla humano consiste en:

- *Pulmones*: fuente de aire durante el habla.
- *Cuerdas vocales* (laringe): responsables de la cualidad de sonoridad de los sonidos sonoros.

  Suponiendo que las cuerdas vocales están cerradas durante, en la producción de un *sonido sonoro*, se incrementa la presión lo suficiente para forzar que las cuerdas vocales se separen. Al separarse, el aire pasa a través de ellas y la presión disminuye, momento en el que la fuerza de los músculos hace que las cuerdas vocales vuelven a juntarse. Cuando las cuerdas vocales vuelven a juntarse, el flujo de aire disminuye, lo que provoca de nuevo un aumento de la presión. Esta vibración de las cuerdas vocales produce pulsos casi periódicos de aire que excitan el sistema por encima de la laringe. A esta frecuencia de vibración se denomina frecuencia fundamental del sonido.

  Cuando las cerdas vocales están demasiado flojas o tensas para vibrar periódicamente, el sonido es *sordo*.
- *Velo del paladar (paladar suave)*: opera como una válvula que permite pasar el aire (y por lo tanto resonar) a través de la cavidad nasal. Los sonidos producidos cuando está abierto incluye la *m* y *n*.
- *Paladar duro*: una larga superficie relativamente dura que constituye el techo de la boca, la cual, permite la articulación de consonantes cuando la lengua se encuentra contra el.

- *Lengua*: articulador flexible, que se mantiene separado del paladar para la pronunciación de las vocales, y entra en contacto con el paladar o otra superficie dura para la articulación de consonantes. La raíz de la lengua forma la pared frontal de la faringe, y sus movimientos le permiten modificar la Sección de la cavidad bucal (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice o la parte trasera con alguna zona del paladar.
- *Dientes*: otro lugar de articulación usado por la lengua como pared en la articulación de ciertas consonantes.
- *Labios*: con movimientos de protuberancia que afectan a la calidad de las vocales, de apertura y de cierre. Completamente cerrados impiden el pasa del flujo de aire oral para la articulación de ciertas consonantes (p, b, m).

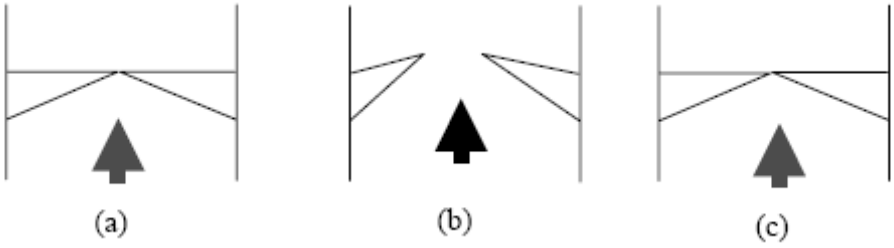

Figura 8- Ciclo de las cuerdas vocales en la laringe [Huang *et al.*, 2001; p. 26]. (a) Cuerdas vocales cerradas (incremento de la presión subglotal). (b) Apertura de las cuerdas vocales debido a la alta presión subglotal alcanzada (disminución de la presión subglotal). (c) La fuerza de los músculos cierran las cuerdas vocales como consecuencia de la disminución de la presión (comienzo del próximo ciclo).

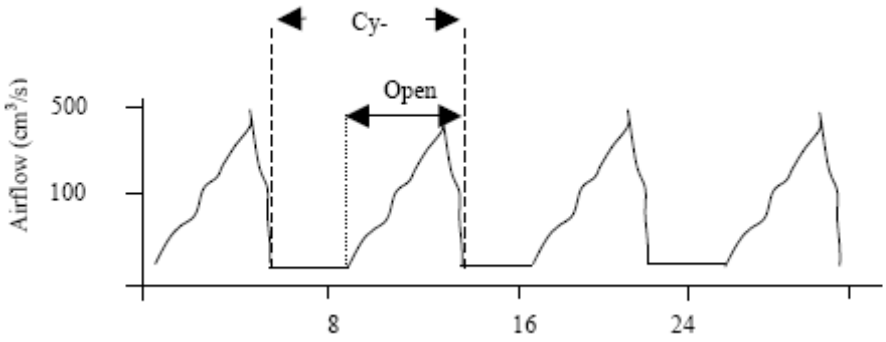

Figura 9- Flujo de aire durante el ciclo de la laringe [Huang *et al.*, 2001; p. 27].

2.2.4. Percepción del habla.

El sistema de percepción auditivo está compuesto por los órganos auditivos periféricos (oído) y el sistema nervioso auditivo (cerebro). El oído procesa una señal acústica de presión transformándola en un patrón de vibración mecánico sobre la membrana basilar, representando así el patrón como una serie de pulsos a ser transmitidos por el nervio auditivo. La información percibida es extraída en varias fases del sistema nervioso auditivo.

Como se puede observar en la *Figura 10*, el oído humano se divide en 3 partes:

- *Oído externo*: canaliza la energía acústica y consiste de la parte externa visible y el canal aditivo externo, de aproximadamente 2.5 cm, a través del cual viaja el sonido.
- *Oído medio*: transforma la energía acústica en energía mecánica, transmitiéndola - y amplificándola- hasta el oído interno.
- *Oído interno*: donde se realiza la definitiva transformación de la energía mecánica en impulsos eléctricos.

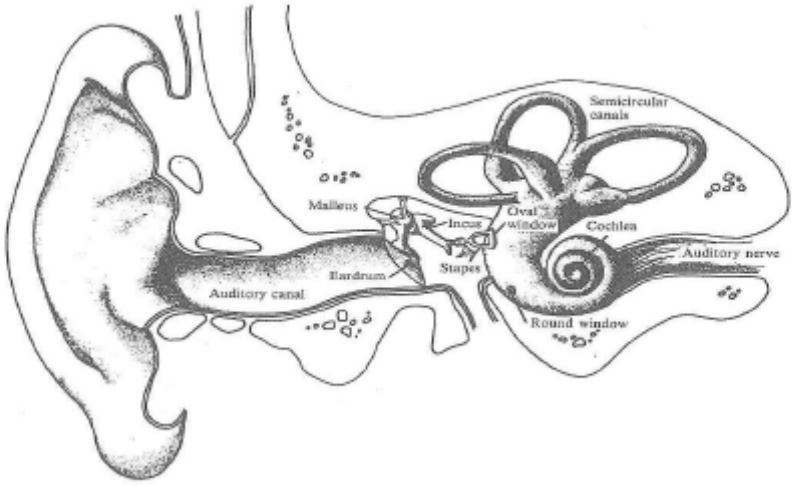

Figura 10- Estructura del sistema periférico auditivo [Huang *et al.*, 2001; p. 29].

Cuando el sonido llega al oído, las ondas sonoras son recogidas por el *pabellón auricular* (o *aurícula*). El pabellón auricular, por su forma helicoidal, funciona como una especie de "embudo" que ayuda a dirigir el sonido hacia el interior del oído. Sin la existencia del pabellón auricular, los frentes de onda llegarían de forma perpendicularmente y el proceso de audición resultaría ineficaz (gran parte del sonido se perdería):

- Parte de la vibración penetraría en el oído.
- Parte de la vibración rebotaría sobre la cabeza y volvería en la dirección de la que procedía. *(reflexión)*.
- Parte de la vibración lograría rodear la cabeza y continuar su camino. (*difracción*).

El pabellón auricular humano es mucho menos direccional que el de otros animales (como los perros) que poseen un control voluntario de su orientación.
Una vez que ha sido recogido el sonido, las vibraciones provocadas por la variación de presión del aire cruzan el canal auditivo externo y llegan a la membrana del tímpano, ya en el oído medio.

El conducto auditivo actúa como una etapa de potencia natural que amplifica automáticamente los sonidos más bajos que proceden del exterior. Al mismo tiempo, en el caso contrario, si se produce un sonido muy intenso que puede dañar el oído, el conducto auditivo segrega cerumen (cera), con lo que cierra parcialmente el conducto, protegiéndolo.

En el oído medio, se produce la transducción, es decir, la transformación la energía acústica en energía mecánica. En este sentido, el oído medio es un transductor mecánico-acústico. Además de transformar la señal, antes de que ésta llegue al oído interno, el oído medio la habrá amplificado.

La presión de las ondas sonoras hace que el tímpano vibre empujando a los osículos, que, a su vez, transmiten el movimiento del tímpano al oído interno. Cada osículo empuja a su adyacente y, finalmente a través de la ventana oval. Es un proceso mecánico, el pie del *estribo* empuja a la *ventana oval*, ya en el oído interno. Esta fuerza empuja a la venta oval es unas 20 veces mayor que la que empujaba a la membrana del tímpano, lo que se debe a la diferencia de tamaño entre ambas.

Esta presión ejercida sobre la ventana oval, penetra en el interior de la *cóclea*, la cual se comunica directamente con el nervio auditivo, conduciendo una representación del sonido al cerebro. La cóclea es un tubo en forma de espiral (de 3.5 cm aproximadamente). La espiral es dividida longitudinalmente por la membrana basilar en dos cámaras que contienen líquido linfático.

2.2.4.1. Análisis espectral.

La cóclea puede ser aproximada como un *banco de filtros*. Los filtros correspondientes al extremo más próximo a la ventana oval y al tímpano responden a las altas frecuencias, ya que la membrana es rígida y ligera. Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, por lo que los filtros correspondientes responden a las bajas frecuencias. Por ello los investigadores emprenden trabajo psicoacústicos experimentales para obtener las escalas de frecuencias que modelen la respuesta natural del sistema de percepción humano.

AT&T Bell Labs ha contribuido de manera muy influyente en los descubrimientos en audición, tales como a banda crítica y el índice de articulación [E. Campbell, 1997]. El trabajo de Fletcher [M. Helander, 1997] apunta a la existencia de bandas críticas en la respuesta de la cóclea. Una banda crítica constituye el intervalo de frecuencia en el cual el oído interno efectúa una integración espacial de la intensidad de la señal sonora, y son de gran importancia para entender fenómenos de adición tales como la percepción de la intensidad, del tono y del timbre.

El sistema de audición lleva a cabo un análisis espectral de sonidos dentro de sus componentes de frecuencia. La cóclea actúa como si estuviese compuesta de filtros superpuestos con un ancho de banda igual al ancho de banda crítico.

Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal, existen diferentes escalas. Una clase de escala de banda crítica es la llamada *escala de frecuencia Bark*. El rango de la escala Bark es de 1 a 24 Barks, correspondientes a 24 bandas críticas del oído. Como muestra la *Figura 11*, la resolución perceptiva es mejor para las bajas frecuencias. Nótese que las bandas críticas del oído son continuas, y un tono de cualquier frecuencia audible siempre encuentra una banda crítica centrado en dicha frecuencia. La frecuencia Bark *b* puede ser expresada en términos de la frecuencia (en Hz) como

$$b(f) = 13 * \arctan(0.00076 * f) + 3.5 * \arctan\left((f / 7500)^2\right) \quad \textit{(Bark)}$$

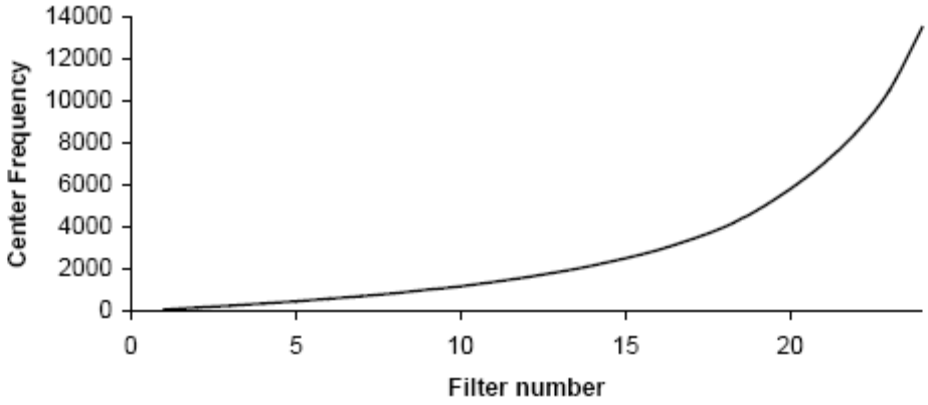

Figura 11- Frecuencia central de los 24 filtros Bark [Huang *et al.*, 2001; p. 33].

Otra importante escala es la *escala Mel*, que ha sido ampliamente utilizada en modernos sistemas de reconocimiento de habla. Pede ser aproximada como

$$B(f) = 1125 * \ln(1 + f / 700)$$

2.3. Fonética y fonología.

2.3.1. Introducción.

Los fonemas son unidades teóricas, postuladas para estudiar el nivel fonético-fonológico de una lengua humana. Entre los criterios para decidir, qué constituye o no un fonema se requiere que exista una función distintiva: son sonidos del habla que permiten distinguir palabras en una lengua. Así, los sonidos /t/ y /d/ son fonemas del español porque existen palabras como /pata/ y /bata/ que tienen significado distinto y su pronunciación sólo difiere en relación con esos dos sonidos (sin embargo en chino los sonidos [p] y [b] son percibidos como variantes posiciónales del mismo fonema). Esto se puede estudiar con más profundidad en Gil, Juana (1989), Llisterri, Joaquim (1991) y Trubetzkoy, Nikolai S. (1939).

Desde un punto de vista estructural, el fonema pertenece a la lengua, mientras que el sonido pertenece al habla. La palabra <casa>, por ejemplo, consta de cuatro fonemas (/k/, /a/, /s/, /a/). A esta misma palabra también corresponden en el habla, acto concreto, cuatro sonidos, a los que la fonología denominará alófonos, y estos últimos pueden variar según el sujeto que lo pronuncie. La distinción fundamental de los conceptos fonema y alófono, está en que el primero es una huella psíquica de la neutralización de los segundos que se efectúan en el habla.

2.3.2. Fono y fonema.

Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. En este sentido, un fonema puede ser representado por una familia o clase de equivalencia de sonidos (técnicamente denominados fonos), que los hablantes asocian a un sonido específico durante la producción o la percepción del habla. Así por ejemplo, en español el fonema /d/ [+ obstruyente, + alveolar, + sonoro] puede ser articulado como oclusiva [d] a principio de palabra o tras nasal o pausa larga, pero es pronunciado como aproximante [ð] entre vocales o entre vocal y líquida, así /dedo/ se pronuncia [deðo] donde el primer y tercer sonido difieren en el grado de obstrucción aunque son similares en una serie de rasgos (los propios del fonema).

Un sonido o fono se caracteriza por una serie de rasgos fonéticos y articulatorios, el número de dichos rasgos y la identificación de los mismos es tarea de la ***fonética***. Un fono es cualquiera de las posibles realizaciones acústicas de un fonema. Por tanto, la fonética es la rama de la lingüística que estudia la producción y percepción de los sonidos de una lengua en sus manifestaciones *físicas* y sus principales ramas son: fonética experimental, fonética articulatoria, fonemática y fonética acústica.

La ***fonología*** en cambio no necesariamente trata entes claramente distinguibles en términos acústicos. Como realidad mental o abstracta, un fonema no tiene porqué tener todos los rasgos fonéticos especificados. Por ejemplo, en diversas lenguas la aspiración es relevante para distinguir pares mínimos, pero un fonema del español puede pronunciarse más o menos aspirado según el contexto y la variante lingüística del hablante pero en general no está especificado el grado de aspiración. En cambio, en lenguas como el chino mandarín o el coreano un fonema tiene predefinido el rasgo de aspiración.

El número de fonemas de una lengua es finito y limitado en cada lengua al número de alófonos potencialmente definibles, si especificamos rasgos fonéticos muy sutiles, es potencialmente ilimitado y varían según el contexto fonético y la articulación individual de los hablantes. En cuanto al número de fonemas no tiene porqué ser fijo, pudiendo cambiar con el número de especificaciones que se dé para cada fonema. Sin embargo, la mayoría de los análisis del español está en torno a 24 unidades (5 vocales y 19 consonantes), aunque no todas las variedades de español tienen el mismo número de fonemas. En cuanto al inglés americano el conjunto de fonemas usado es mayor, 40 fonemas. En otras lenguas como el ruso que llegan a 48 fonemas.

Dada la distinción entre fonema y fono, existe otra forma de concebir un fonema como una especificación incompleta de rasgos fonéticos. Esta relación es de hecho equivalente a la del fonema como conjunto de fonos: el fonema sería el conjunto de rasgos fonéticos comunes a todos los fonos que forman la clase de equivalencia del fonema.

Fijado un conjunto de rasgos fonéticos se pueden definir los sonidos de la lengua. En principio no hay límite a lo fina que pueda ser la distinción que establecen estos rasgos. Potencialmente la lista de sonidos puede hacerse tan grande como se quiera si se incluyen más y más rasgos. Sin embargo el número de fonemas es un asunto diferente, puesto que muchos de los anteriores sonidos serán equivalentes desde el punto de vista lingüístico. Un sistema fonológico es un par $\mathcal{F} = (F, (R))$, donde Fes un inventario de fonemas abstractos definidos por unos pocos rasgos del conjunto total (las lenguas naturales oscilan

entre 1 o 2 decenas hasta 4 o 5 decenas de fonemas), y $\mathcal{R}$ es el conjunto de reglas que en función del contexto relativo de aparición de los fonemas definen totalmente los rasgos fonéticos, así el conjunto de reglas puede pensarse como una aplicación del conjunto de secuencias admisibles de fonemas al conjunto de secuencias admisibles de sonidos:

$$\mathcal{R} : P_0(F) \rightarrow P_0(S)$$

, donde $P_0(F), P_0(S)$ representan el conjunto de secuencias finitas de fonemas y el conjunto de secuencias finitas de sonidos.

2.3.2.1. Características de un fonema.

Podemos decir que fonema es una unidad fonológica diferenciadora, indivisible y abstracta.

- **Diferenciadora:** porque cada fonema se delimita dentro del sistema por las cualidades que se distinguen de los demás y además es portador de una significativa especial. Por ejemplo, /k-o-t-a/ y /b-o-t-a/ son dos palabras que se distinguen semánticamente debido a que /k/ se opone a /b/ por la sonoridad.
- **Indivisible**: no se puede descomponer en unidades menores. Por ejemplo, la sílaba o el grupo fónico sí pueden fraccionarse. Un análisis pormenorizado del fonema revela que está compuesto por un haz de diversos elementos fónicos llamados rasgos distintivos, cuya combinación forma el inventario de fonemas. El inventario de rasgos distintivos es asimismo limitado y viene a constituir una especie de tercera articulación del lenguaje.
- **Abstracta**: no son sonidos, sino modelos o tipos ideales de sonidos. La distinción entre sonido y fonema ha sido un gran logro en los últimos tiempos en la lingüística.

2.4. Creación de modelos fonéticos.

La creación de los modelos acústicos, para su posterior uso en reconocimiento de habla, consiste principalmente en 2 etapas:

- ***Asignación*** de los parámetros extraídos a las representaciones discretas de nuestro diseño (fonemas, tri-fonemas, palabras…) correspondientes, con el objetivo de crear un modelo para cada una de las representaciones discretas que las identifique. Tanto las técnicas para la extracción de parámetros como los distintos modelos, son definidos en las siguientes Secciones.

- ***Entrenamiento*** de los modelos acústicos creados para cada una de las distintas representaciones discretas definidas. Para dotar de robustez a los modelos es necesario una gran cantidad de ficheros de voz con sus respectivas transcripciones.

Los modelos acústicos obtenidos dan lugar a un módulo conocido como decodificador acústico-fonético, encargado del reconocimiento del habla. La entrada al módulo será, por tanto, la señal a decodificar, que será sometida al mismo proceso de parametrización que los datos de entrenamiento.

Existen varias técnicas para el modelado acústico. Algunas de ellas son:

- ***Hidden Markov Models (HMMs)***: máquina de estados finitos, en la que las observaciones son una función probabilística del estado, es decir, el modelo es un proceso doblemente estocástico.

- ***DTW (Alineamiento temporal Dinámico)***: consiste en alinear de forma temporal los parámetros del archivo de test y los parámetros de los modelos, obteniendo la función que alinea a ambos, eligiendo la función de menor coste posible para dicha adaptación. En la F*igura 12* se muestra como representar la función de adaptación.

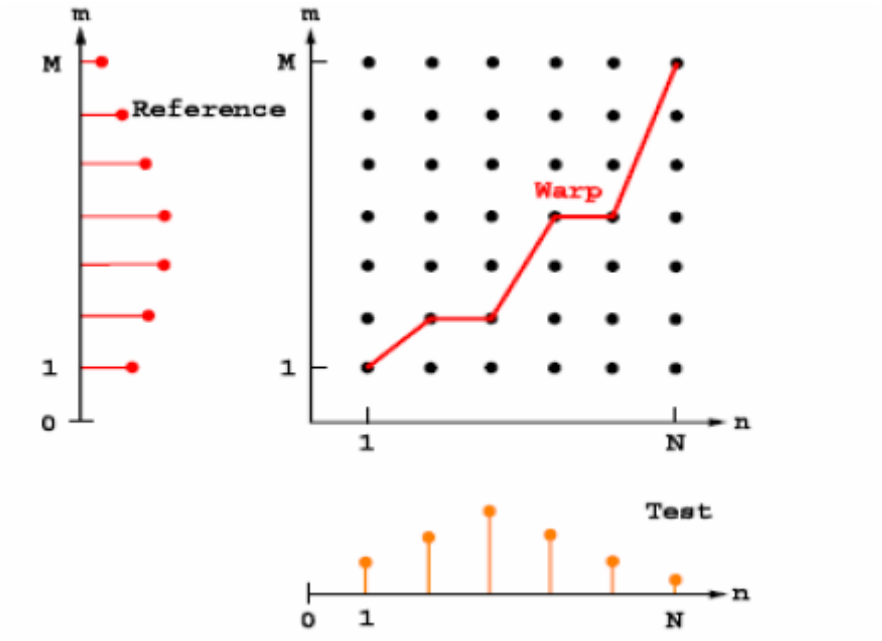

Figura 12- Función de adaptación DTW.

- **VQ** (Cuantificación vectorial) [R.O. Duda, 2001] consiste en representar los parámetros obtenidos de los fonemas como un espacio vectorial de dimensión el número de parámetros. De esta forma, al fonema a reconocer se le asigna el vector, cuya distancia a él sea mínima. Así, los fonemas quedan representados por unos vectores, denominados (centroides).Por tanto, a todos los puntos de una zona determinada se les asigna el vector correspondiente. En la *Figura 1* se muestra un espacio bidimensional,(el número de parámetros usado son dos) en el que los puntos verdes son los vectores de test, mientras que los rojos son los vectores a los que se asignan (obtenidos de forma óptima durante el entrenamiento), siendo cada una de las regiones los fonemas posibles.

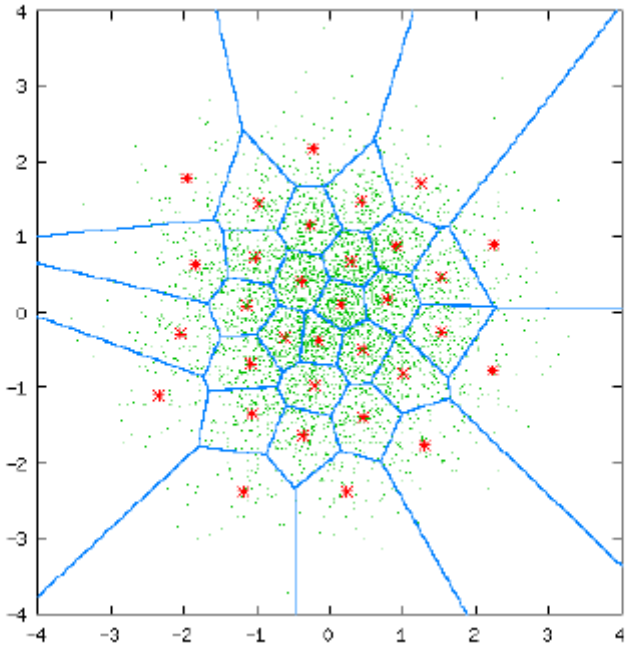

Figura 13- VQ bidimensional.

Este sistema tiene los siguientes beneficios:

- ✓ Reducción notable de la capacidad de almacenamiento necesaria para la obtención de la información del análisis espectral.
- ✓ Reducción de la complejidad en el cómputo de las distancias.
- ✓ Representación discreta de los "sonidos" de voz.

Por el contrario presenta las siguientes desventajas:

- × Introducción de distorsión espectral, como consecuencia del error de cuantificación que se produce al asignar a un fonema (con un vector de parámetros determinado) uno de los vectores (centroides) obtenidos en el entrenamiento. Este error de cuantificación se puede disminuir, aumentando el número de centroides (aumento del *codebook*)
- × Aumento de la complejidad en la asignación del centroide óptimo con el aumento del *codebook*.
- × Problemas de almacenamiento según aumentamos el *codebook*.

2.4.1. Hidden Markov Models (HMMs).

2.4.1.1. Introducción.

La popularidad de los métodos estadísticos de cadenas de Markov o modelos ocultos de Markov (HMMs) ha sido incrementada en los últimos años. Existen 2 fuertes razones por lo que esto ha ocurrido. En primer lugar, los modelos son muy ricos en estructuras matemáticas y por lo tanto se puede formar las bases teóricas para su uso en un amplio rango de aplicaciones. En segundo lugar, los modelos, cuando son propiamente aplicados, trabajan muy bien para varias aplicaciones importantes.

Ni la teoría de los HMMs ni sus aplicaciones en reconocimiento de habla son nuevas. La teoría básica fue introducida en una serie de clásicas publicaciones por Baum y sus colaboradores [L. E. Baum, T. Petrie, 1966, pp. 1554-1563; L. E. Baum, J. A. Egon, 1967, pp. 360-363; L. E. Baum, J. R. Sell, 1968, pp. 211-227; L. E. Baum *et al.*, 1970, pp. 164-171; L. E. Baum, 1972, pp. 1-8] al final de los 1960s y principios de los 1970s y fue implementado para aplicaciones en reconocimiento de habla por Baker [J. K. Baker, 1975,pp. 24-29] en CMU y por Jelinek [F. Jelinek, 1969, pp. 675-685; L. R. Bahl, F. Jelinek, 1975, pp. 404-411; F. Jelinek *et al.*, 1975, pp. 250-256; F. Jelinek, 1976, pp. 532-536; R. Bakis, 1976; F. Jelinek *et al.*, 1982; L. R. Bahl *et al.*, 1983, pp. 179-190] y sus colaborados en IBM. Sin embargo, el general entendimiento y aplicaciones de la teoría de los HMMs al reconocimiento de habla han aparecido en los últimos años. Esto es debido a que la teoría básica de los HMMs fue publicada en revistas matemáticas, las cuales, generalmente, no son leídas por ingenieros que trabajan en problemas de reconocimiento del habla [R. R. Lawrence, Fellow, 1989, pp. 257-286].

Un HMM es una máquina de estados finita, en la que las observaciones son una función probabilística del estado, es decir, el modelo es un proceso doblemente estocástico formado por un proceso estocástico oculto no observable directamente, que corresponde a las transiciones entre estados y un proceso estocástico observable cuya salida es la secuencia de vectores espectrales.

En un modelo observable de Markov, el estado es lo que es directamente visible, por lo tanto, los únicos parámetros que existen son las probabilidades de transición entre estados. Por el contrario, en un HMM el estado no es visible directamente sino que sólo son visibles las variables influenciadas por el estado. Cada estado tiene una distribución de probabilidad sobre el símbolo a la salida. En nuestro caso, la variable observable la consideramos continua, por tanto, empleamos una función de densidad de probabilidad continua modelada como una mezcla de Gaussinas. Esto es lo que se denomina *Continuous–density HMM o CDHMM.*

Un ejemplo de cadena de Markov observable no relacionado con el habla pero que explica las diferencias entre un HMM y un Modelo de Markov observable es el siguiente:

La cotización en el *DowJones*, en la que cada estado marca si la cotización ha subido, bajado o no ha cambiado con respecto al día anterior, la representación de este modelo sería como sigue:

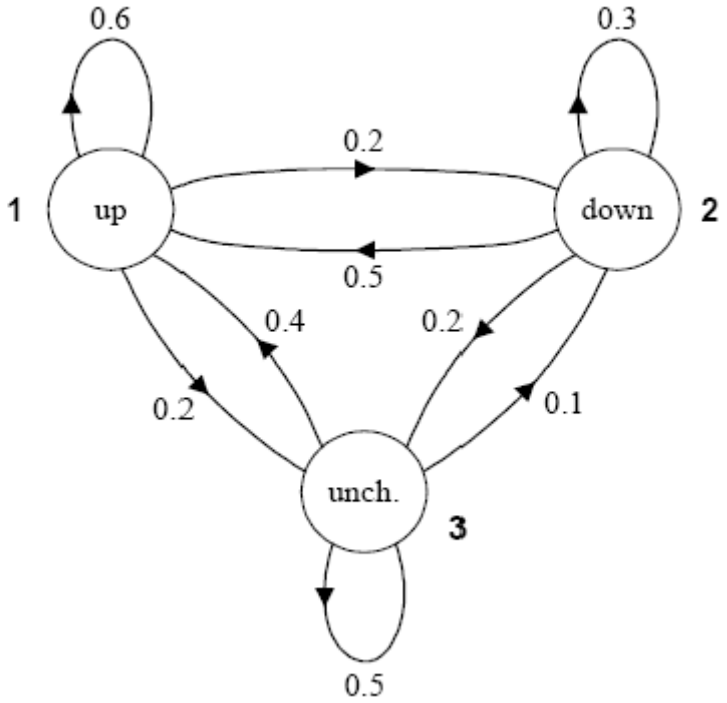

Figura 14- Esquema de una cadena de Markov para la cotización en el *DowJones* [Huang *et al.*, 2001; p. 381].

A este gráfico habría que añadir las probabilidades de comienzo en cada uno de los estados. Como se puede ver en este esquema la salida observable es determinista para cada estado, por tanto se sabe en que estado del modelo se está en ese momento.

Por el contrario un modelo de HMM, también del *DowJones*, nos daría el esquema mostrado en la F*igura 15*, en el que ahora cada uno de los estados es un mercado concreto, mientras que las salidas son si suben o bajan o no cambian. Es decir, ahora la salida de cada estado se asigna mediante un proceso estocástico.

La técnica de HMM se usa en la actualidad en aquellos sistemas en los que el modelado tiene una dependencia del tiempo, como pueden ser los sistemas reconocimiento fonético y del habla en general.

Una razón, por la que los HMMs son utilizados en el reconocimiento de fonemas, es que una señal de voz puede verse como una señal invariante a corto plazo (de unos 10 -20 milisegundos). La voz se podría interpretar así como un modelo de Markov para muchos procesos estocásticos (conocidos como **estados**). Otra razón por la que los HMMs son populares, es que pueden ser entrenados automáticamente, siendo factible realizar los cálculos en un tiempo razonable.

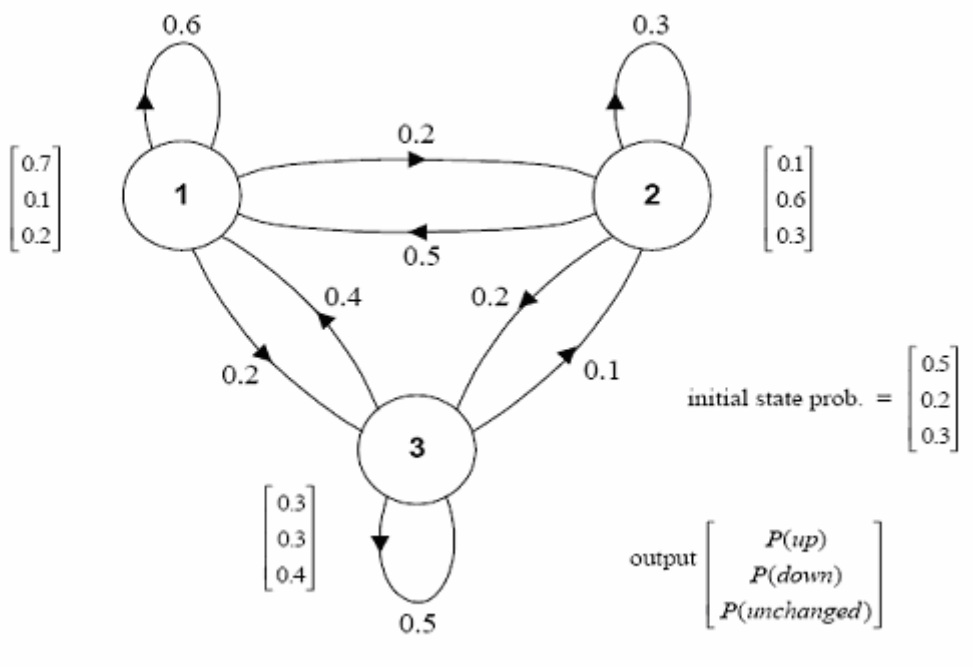

Figura 15- Esquema de un HMM para la cotización en el Dow Jones [Huang *et al.*, 2001; p. 381].

El modelo oculto de Markov tendrá en cada estado una distribución estadística llamada mezcla de Gaussianas de matriz de covarianza diagonal, que da una probabilidad para cada vector observado. Por tanto, cada fonema tendrá una distribución de salida. Un modelo oculto de Markov para una secuencia de fonemas se construye concatenando los modelos ocultos entrenados para los fonemas separados.

El uso de los HMM permite eludir las limitaciones de algunos otros sistemas en el reconocimiento de fonemas como:

- DTW (Alineamiento temporal Dinámico), en el que no hay posibilidad de realizar un entrenamiento estadístico, ya que se realiza comparaciones entre secuencias de vectores de parámetros.

- VQ (Cuantificación temporal) en que se hace una asignación dura entre los vectores y la clase que modela. Además tiene que respetar el compromiso entre el tamaño del *codebook* y el error de cuantificación.

2.4.1.2. Elementos de un HMM.

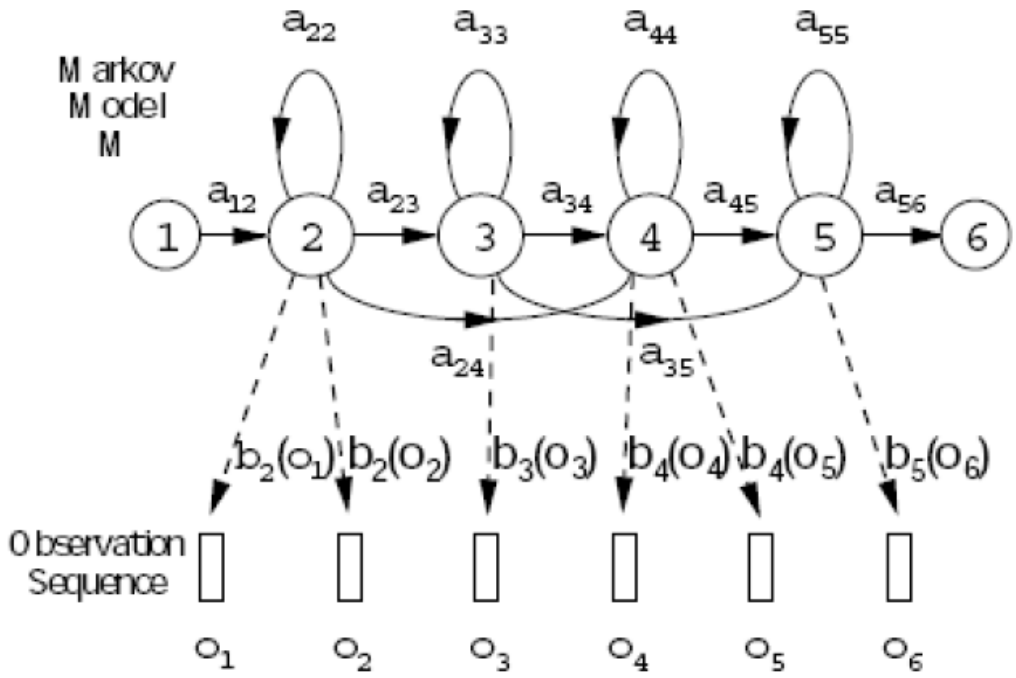

Figura 16- El modelo de generación de Markov[The HTK book, 2005].

Suponiendo un HMM discreto, en que las observaciones posibles pertenecen a un conjunto discreto, entonces el HMM (véase *Figura 16*) vendrá dado por:

- *N*: el número de estados del modelo, donde q_t denota el estado en el instante de tiempo t. Los HMMs que vamos a utilizar están compuestos por 5 estados. Sin embargo en HTK, tanto el estado 1 como el estado 5 no generan ninguna salida.

$S = \{s_1, s_2, ..., s_N\}$

- La dimensión del conjunto de observaciones distintas de salida *M*, es decir el tamaño del alfabeto

$V = \{v_1, v_2, ..., v_M\}$

- La distribución de probabilidad de transición entre estados $A=\{a_{ij}\}$:

$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i)$, $1 \leq i,j \leq N$

- La distribución de probabilidades de emisión de símbolos entre estados

$B = \{b_j(k)\}$

$B_j(O_k) = P(O_k \mid q_t = s_j)$, $1 \leq j \leq N$, $1 \leq k \leq M$ y donde O_k es un símbolo perteneciente a V.

- Distribución del estado inicial $\pi = \{\pi_i\}$:

$\pi_i = P(q_0 = s_i)$, $1 \le i \le N$

Con todo esto, un HMM se describe como $\lambda = \{A, B, \pi\}$.

Dada esta definición, surgen tres problemas que es necesario resolver para que los HMMs tengan utilidad en aplicaciones reales:

1. Problema de evaluación de la probabilidad.
2. Problema de encontrar la secuencia de estados óptima (Problema de decodificación).
3. El problema de entrenamiento de un modelo (Problema de aprendizaje).

2.4.1.3. Problemas a resolver para la utilización de un HMM.

Dada la definición de la Sección anterior para un HMM, existen tres problemas básicos a resolver para que el modelo pueda se usado en aplicaciones reales:

1. ***Problema de evaluación de la probabilidad:*** dada una secuencia de observaciones $O = \{o_1, o_2, ..., o_T\}$ (siendo T la longitud de la secuencia de observación) y el modelo $\lambda = \{A, B, \pi\}$, el problema es cómo obtener de forma eficiente $P(O \mid \lambda)$. Es decir, la probabilidad de obtener una secuencia de observación dado un modelo determinado. La resolución de este problema nos permite la evaluar cómo es de bueno el modelo HMM para una secuencia de observación.

2. ***Problema de decodificación***: dada una secuencia de observaciones $O = \{o_1, o_2, ..., o_T\}$ y el modelo $\lambda = \{A, B, \pi\}$, encontrar la secuencia de estados $Q = \{q_1, q_2, ..., q_T\}$ más probable, para la secuencia de observaciones dada. Al solventar este problema podemos obtener la decencia de estados óptima para una decencia de observaciones.

3. ***El problema de aprendizaje***: Como maximizar los parámetros del modelo H ! para obtener la máxima $P(O \mid \lambda)$ para unas observaciones de entrenamiento O. Este problema es crucial para la mayoría de aplicaciones de los HMMs, ya que nos permite adaptar de forma óptima los parámetros del modelo a los datos de entrenamiento observado.

- Evaluación de un HMM- Algoritmo forward-backward.

La manera más intuitiva de solucionar el problema de evaluación sería enumerando todas las posibles secuencias de estados de longitud T que generen la secuencia de observación O y sumando sus probabilidades según el teorema de la Probabilidad Total:

$$P(O \mid \lambda) = \sum_{Q} P(O \mid Q, \lambda) \cdot P(Q \mid \lambda) \qquad (1)$$

Para ello consideremos una determinada secuencia de estados: $Q = \{q_1, q_2, ..., q_T\}$ donde q_1 es el estado inicial. La probabilidad de la secuencia de observación O dada la secuencia de estados Q es:

$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O_t \mid q_t, \lambda)$, donde se asume independencia estadística de las observaciones. Por lo tanto se obtiene:

$$P(O \mid Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdot ... \cdot b_{q_T} .$$

Por otra parte la probabilidad de la secuencia de estados Q se puede expresar como:

$P(Q \mid \lambda) = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdot ... \cdot a_{q_{T-1} q_T}$, que se interpreta como la probabilidad del estado inicial, multiplicada por las probabilidades de transición de un estado a otro.

Sustituyendo los dos términos anteriores en el sumatorio inicial (Ecuación 1) se obtiene la probabilidad de la secuencia de observación:

$$P(O \mid \lambda) = \sum_{Q} P(O \mid Q, \lambda) \cdot P(Q \mid \lambda) = \sum_{q_1, q_2, ..., q_T} \pi_{q_1} \cdot b_{q_1}(O_1) \cdot a_{q_1 q_2} \cdot b_{q_2}(O_2) \cdot ... \cdot a_{q_{T-1} q_T} \cdot b_{q_T}(O_T)$$

La interpretación del resultado obtenido es la siguiente: Inicialmente en el tiempo $t = 1$ nos encontramos en el estado q_1 con probabilidad π_{q_1} y generamos el símbolo O_1 con probabilidad $b_{q_1}(O_1)$. Al avanzar el reloj al instante $t = 2$ se produce una transición al estado q_2 con probabilidad $a_{q_1 q_2}$ y generamos el símbolo O_2 con probabilidad $b_{q_2}(O_2)$.
Este proceso se repite hasta que se produce la última transición del estado q_{T-1} al estado q_T con probabilidad $a_{q_{T-1} q_T}$ y generamos el símbolo O_T con probabilidad $b_{q_T}(O_T)$.

A pesar de haber llegado al resultado deseado se puede ver fácilmente que no es una manera muy eficiente de calcular la probabilidad ya que requiere realizar $2 \cdot T \cdot N^T$ operaciones, ya que para cada T se pueden alcanzar N^T posibles secuencias de estados, lo que resulta computacionalmente intratable.

Afortunadamente existe una manera más eficiente de llegar al mismo resultado. La clave está en guardar los resultados intermedios y utilizarlos para los posteriores cálculos de la secuencia de estados. A este algoritmo se le denomina el *Algoritmo de Avance (Forward)*.

El primer paso es definir la variable hacia delante como $\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = S_i \mid \lambda)$. Esta variable corresponde con la probabilidad de que el modelo λ se encuentre en el estado *i*, habiendo generado la secuencia parcial $O_1, O_2, ..., O_t$ hasta el instante de tiempo t. $\alpha_t(i)$ se puede calcular por inducción siguiendo los siguientes pasos:

1. Inicialización:

$\alpha_1(i) = \pi_i \cdot b_i(O_1)$, $1 \le i \le N$

En este paso se inicializan las probabilidades hacia delante como la probabilidad conjunta del estado S_i y la observación inicial O_1.

2. Inducción:

$$\alpha_{t+1}(j) = \left[\sum_{i=1} \alpha_t(i) \cdot a_{ij} \right] \cdot b_j(O_{t+1}) \ , \ 1 \le t \le T\text{-}1, \ 1 \le j \le N$$

La expresión entre corchetes representa la probabilidad de alcanzar el estado S_j en el instante de tiempo $t+1$ partiendo de todos los estados posibles S_i en el instante *t* habiendo observado hasta el instante *t* la secuencia parcial $O_1, O_2, ..., O_T$. Si multiplicamos ahora dicho término por la probabilidad de observar O_{t+1} se obtiene $\alpha_{t+1}(j)$.

3. Finalización:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

El cálculo de $P(O \mid \lambda)$ final se realiza sumando todas las variables hacia delante $\alpha_T(i)$ en el instante final *T*. Esto es así, ya que por definición $\alpha_T(i)$ es igual a la probabilidad conjunta de haber observado la secuencia $O_1, O_2, ..., O_T$ y encontrarnos en el estado S_i: $\alpha_T(i) = P(O_1, O_2, ..., O_T, q_T = S_i \mid \lambda)$, con lo que si sumamos dicha probabilidad para todos los estados posibles obtenemos la probabilidad esperada $P(O \mid \lambda)$.

La complejidad de este algoritmo comparado con la manera directa de calcular $P(O \mid \lambda)$ es mucho menor y se encuentra en el orden de $O(N^2 \cdot T)$, con lo que se el ahorro computacional es claro.

Otro algoritmo semejante al *forward* es el *backward* que consiste en lo siguiente:

La probabilidad de observación de una secuencia en el estado *i* y con un modelo λ es:

$\beta_t(i) = P(O_{t+1}, O_{t+2}, ..., O_T \mid q_t = S_i, \lambda)$, donde $\beta_t(i)$ es la probabilidad de generar una secuencia de observación parcial *t+1* hasta el final, dado el estado S_i en *t* y el modelo λ.

Podemos resolver $\beta_t(i)$ de forma inductiva:

1. Inicialización:

$$\beta_t(i) = \frac{1}{N} \quad , \ 1 \le t \le N$$

Todos los estados son equiprobables.

2. Inducción:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \ , \ \text{t= T-1, T-2,… ,1} \ , \ 1 \le \text{i} \le \text{N}$$

La relación entre α y β adyacentes se puede observar mejor en la *Figura 17*. α se calcula recursivamente de izquierda a derecha mientras β se calcula recursivamente de derecha a izquierda.

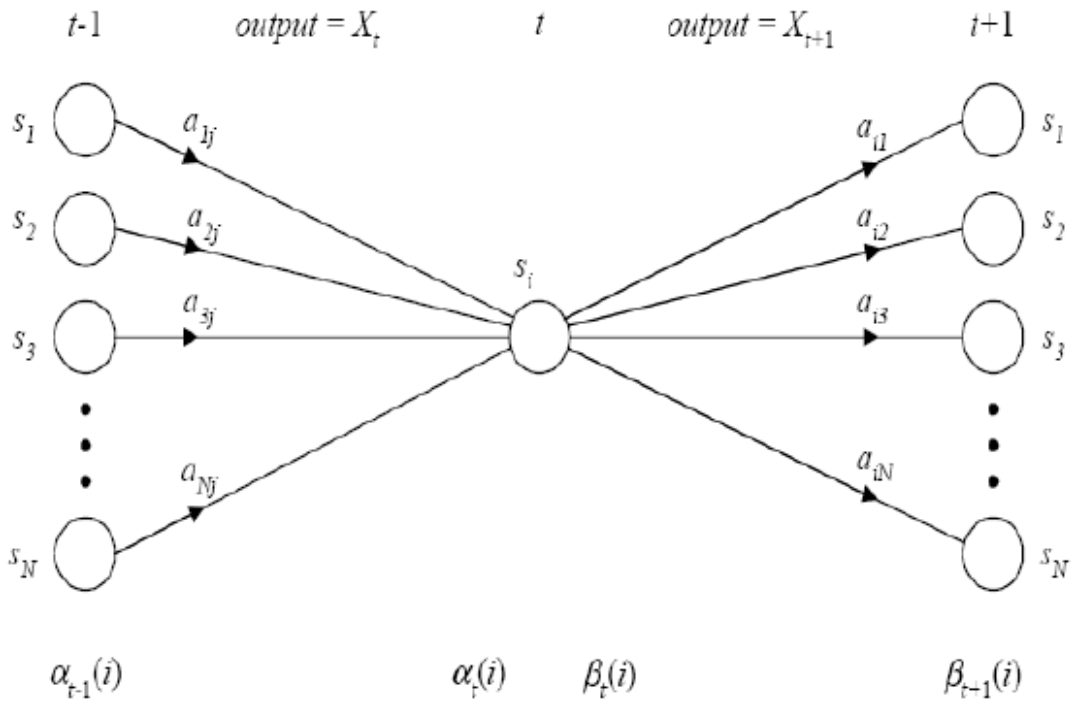

Figura 17- Relación de α_{t-1} y α_t y de β_t y β_{t+1} en el *algoritmo forward-backward* [Huang *et al.*, 2001; p. 390].

- Decodificación de n HMM – Algoritmo de Viterbi.

Decodificar un HMM consiste en encontrar la secuencia de estados óptima, dada una secuencia de observación. La resolución de este problema resulta muy importante para tareas de segmentación y reconocimiento de voz.

A diferencia del problema de evaluación, para el que se puede dar una solución exacta, existen diferentes maneras de resolver este problema. La razón es que la definición de secuencia óptima no es única, sino que existen varios criterios de optimización.

Un criterio de optimización podría ser seleccionar aquellos estados que tengan individualmente la probabilidad más alta de ocurrencia. Sin embargo, este método no parece el más acertado, ya que no tiene en cuenta la probabilidad de ocurrencia de secuencias de estados. Por ejemplo, la probabilidad de transición entre determinados estados es cero ($a_{ij} = 0$), este criterio nos podría dar como solución al problema una secuencia de estados que no fuera válida.

Este problema puede resolverse con el ***algoritmo de Viterbi***, que es similar al algoritmo anterior (*Forward*), con la excepción, de que en vez de tomar la suma de valores de probabilidad en los estados anteriores, se toma el máximo de las probabilidades. De esta forma se consigue no solo dar la secuencia de observación más probable sino el camino de

máxima probabilidad, consiguiendo la secuencia de estados que da una mayor probabilidad.

El *algoritmo de Viterbi* es el criterio más extendido. Trata de encontrar la mejor secuencia de estados, es decir, maximizar la probabilidad $P(q \mid O, \lambda)$ o lo que es equivalente, maximizar $P(O, q \mid \lambda)$. En la práctica este método también se puede utilizar para evaluar HMMs.

Para encontrar la mejor secuencia de estados $Q = q_1, q_2, ..., q_T$ para una decencia de observación dada $O = O_1, O_2, ..., O_T$, definimos la variable

$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P[q_1 q_2 ... q_t = i, O_1 O_2 ... O_t \mid \lambda]$, que representa la secuencia de estados con mayor probabilidad en el instante *t* que acaba en el estado S_i y que ha generado las *t* primeras observaciones.

A continuación se sigue n proceso de inducción similar al *Algoritmo Forward-Backward*:

1. Inicialización:

   $\delta_1(i) = \pi_i \cdot b_i(O_1), 1 \le i \le N$

   $\phi_1(i) = 0$

   Inicialmente se define la probabilidad $\delta_1(i)$ como la probabilidad de encontrarse en el estado S_i en el instante $t = 1$ multiplicada por la probabilidad de generar el símbolo O_1.

   El vector ϕ, en el que se va a almacenar el argumento que maximiza $\delta_t(j)$ para cada valor de *t* y de *j*, toma inicialmente el valor *0*.

2. Inducción:

   $\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}] b_j(O_t), 2 \le t \le T, 1 \le j \le N$

   Y se guarda el camino con mayor probabilidad:

   $\phi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}], 2 \le t \le T, 1 \le j \le N$

3. Finalización:

   $P^* = \max_{1 \le i \le N} [\delta_T(i)]$

   $q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$

   Las iteraciones en el punto 3 terminan cuando se han generado las T observaciones.

4. *Backtracking*:

   $q_t^* = \phi_{t+1}(q_{t+1}^*), t = T-1, T-2, ..., 1$

   En este último paso se reconstruye la secuencia de estados, partiendo desde el estado final hasta llegar al principio.

Como se puede observar el algoritmo seguido es muy semejante al de avance hacia delante (Forward) empleado en la fase de evaluación, y el orden de operaciones también está en torno a $O(N^2 \cdot T)$.

- Entrenamiento de un modelo – Algoritmo Baum-Welch.

El último y más complicado de los 3 problemas plantea cómo se deben ajustar los parámetros del modelo $\lambda = \{A, B, \pi\}$ para maximizar la probabilidad $P(O \mid \lambda)$ de la secuencia de observación dado el modelo.

El principal inconveniente es que no existe ningún método analítico conocido que maximice el conjunto de parámetros a partir de los datos de entrenamiento. Se puede resolver, sin embargo, utilizando un procedimiento iterativo como *el algoritmo de Baum-Welch*, también conocido como el algoritmo de avance-retroceso. Este algoritmo usa los mismos principios que el algoritmo EM (*Expectation Maximization*).El procedimiento consiste en actualizar los pesos de forma iterativa para poder explicar mejor las secuencias de entrenamiento observadas.

Un parámetro que debemos definir es el $\xi_t(i, j)$ como la probabilidad de encontrarnos en el estado *i* en el instante *t* y en el estado *j* en el instante *t+1*, para un modelo y una secuencia de observación dados

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j \mid O, \lambda)$$

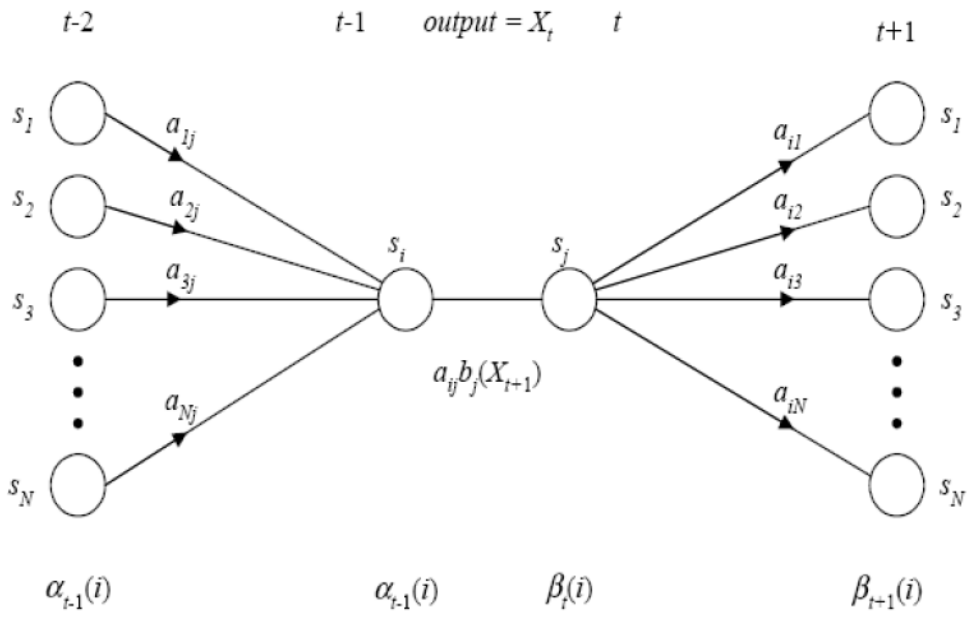

Figura 18- Ilustración de las operaciones necesarias para el cálculo de $\xi_t(i, j)$ [Huang *et al.*, 2001; p. 391].

Utilizando las probabilidades de los métodos forward y backward podemos escribir como sigue:

$$\xi_t(i,j)=\frac{P(q_t=i,q_{t+1}=j\mid O,\lambda)}{P(O\mid\lambda)}=\frac{\alpha_t(i)\cdot a_{ij}\cdot b_j(O_{t+1})\cdot\beta_{t+1}(j)}{P(O\mid\lambda)}=\frac{\alpha_t(i)\cdot a_{ij}\cdot b_j(O_{t+1})\cdot\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\cdot a_{ij}\cdot b_j(O_{t+1})\cdot\beta_{t+1}(j)}$$

Suponiendo $\gamma_t(i)$ la probabilidad de encontrarnos en el estado i en el instante t, para la secuencia de observaciones completa y el modelo dados, podemos calcular $\gamma_t(i)$ a partir de $\xi_t(i,j)$ como sigue:

$$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$$

Si realizamos el sumatorio de $\gamma_t(i)$ sobre el índice t, obtendremos un resultado que puede ser interpretado como el número esperado de veces (en el tiempo) que el estado *i* es visitado o, de manera equivalente, el número esperado de transiciones realizadas desde el estado *i* (excluyendo el instante $t=T$ del sumatorio). De manera similar, el sumatorio de $\xi_t(i,j)$ sobre el índice t (desde $t=1$ hasta $t=T$) puede ser interpretado como el número esperado de transiciones desde el estado *i* al estado *j*. Esto es:

$$\sum_{t=1}^{T-1}\gamma_t(i)=\textit{ Número esperado de transiciones desde el estado } S_i.$$

$$\sum_{t=1}^{T-1}\xi_t(i,j)=\textit{ Número esperado de transiciones desde el estado } S_i \textit{ al } S_j.$$

Usando las formulas anteriores, podemos dar un método para la reestimación de los parámetros de un HMM $\lambda=\{A,B,\pi\}$ como sigue:

$\overline{\pi}_i=$ *Número esperado de veces que permanecemos en el estado* S_i *en el instante* $t=1$, $\gamma_t(i)$.

$$\overline{a}_{ij}=\frac{\textit{número esperado de transiciones desde el estado } S_i \textit{ al estado } S_j}{\textit{número esperado de transiciones desde el estado } S_i}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$

$$\overline{b}_j(k)=\frac{\textit{número esperado de veces en el estado } S_j \textit{ observando el símbolo } v_k}{\textit{número esperado de veces en el estado } S_j}=\frac{\sum_{\substack{t=1\\ s.t.O_t=v_k}}^{T-1}\gamma_t(j)}{\sum_{t=1}^{T-1}\gamma_t(j)}$$

De esta forma, hemos obteniendo un nuevo modelo $\overline{\lambda} = \{\overline{A}, \overline{B}, \overline{\pi}\}$, como resultado de una re-estimación de los parámetros del modelo $\lambda = \{A, B, \pi\}$. Si el modelo λ definía un punto crítico de la función de máxima verosimilitud, tendremos $\overline{\lambda} = \lambda$, o bien el nuevo modelo $\overline{\lambda}$ que cumple $P(O \mid \overline{\lambda}) > P(O \mid \lambda)$. Esto significa que se ha encontrado un nuevo modelo desde el cual la secuencia de observación O es más probable que se haya producido. Por tanto, el *algoritmo de reestimación de Baum-Welch* garantiza una mejora monótona en la probabilidad en cada iteración hasta que ésta converge en un máximo local.

El principal inconveniente del *algoritmo Baum-Welch* es que conduce de forma exclusiva a máximos locales. En la mayoría de los casos de interés la función de verosimilitud es compleja y contiene muchos de estos máximos.

Los modelos que se manejan en este proyecto son *Continuous-Density HMM* con modelos de Gaussianas, por lo que las expresiones enunciadas anteriormente deben ser pasadas al caso continuo.

2.5. Modelado de idioma.

Antes de la década de los 80', los reconocedores automáticos de habla sólo hacían uso de información acústica para evaluar las hipótesis de transcripción. Entonces notaron como, la incorporación de conocimientos a cerca del texto hablado (explotando redundancias textuales), podría mejorar significativamente la precisión del reconocedor automático de habla. El habla usualmente sigue unas reglas lingüísticas (sintácticas y semánticas). Algunas veces, el habla es meramente una secuencia aleatoria de palabras pertenecientes a un vocabulario muy limitado; en tales casos no existe redundancias textuales.

Normalmente, dado un historial de palabras anteriores en una frase, el número P de de palabras que no debe considerar como posible próxima palabra es mucho más pequeño que el tamaño del vocabulario, V. P mide la perplejidad de un modelo de lenguaje (LM), que determina cuanto es de bueno dicho LM. En otras palabras, un modelo de lenguaje, en reconocimiento de habla, intenta predecir la siguiente palabra en una secuencia hablada a partir de las N palabras anteriores. Los LMs son descripciones estocásticas de probabilidades de texto de N palabras consecutivas en los textos de entrenamiento. Los valores típicos de N son 1, 2 (bi-gramas), 3(tri-gramas). Por tanto, en la actualidad es muy común la integración de un LM con los modelos acústicos HMMs.

Típicamente, los modelos de N-gramas estiman la probabilidad de cada palabra, dado el contexto de las N-1 palabras precedentes. Por ejemplo, los modelos de bi-gramas usan estadísticos de parejas de palabras y los modelos de tri.-gramas de tríos de palabras. Para los modelos de uni-gramas las probabilidades obtenidas para cada palabra son independientes del contexto. Estas probabilidades son obtenidas a través del análisis de una gran cantidad de texto, y captura de las redundancias sintácticas y semánticas del texto.

Aunque el principio básico de un LM de N-gramas es my simple, en la práctica hay a menudo más de algún N-grama potencial que siempre puede ser colectado un número suficiente de veces en el texto de entrenamiento para obtener una robusta estimación de las frecuencias. Además, para aplicaciones reales como el reconocimiento de habla, el uso de un texto de entrenamiento estático y finito dificulta la generación de un único LM que sea apropiado para un variado material de texto. Por ejemplo, un LM entrenado con texto

de periódicos, será un buen predictor para noticias dictadas, pero el mismo LM será un pobre predictor para una carta personal o para un interfaz de voz como un sistema de reserva de vuelos. Una última dificultad es que el vocabulario de un LM de N-gramas es finito y fijo en el tiempo de construcción. Entonces, si el LM está basado en palabras, sólo puede predecir palabras que se encentren dentro de dicho vocabulario y además no se peden añadir nuevas palabras sin reconstruir el LM.

2.5.1. Modelos de lenguaje de n-gramas.

Los modelos de lenguaje estiman la probabilidad de una secuencia de palabras, $P(w_1, w_2, ..., w_m)$. Dicha probabilidad puede ser descompuesta como un producto de probabilidades condicionales:

$$P(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, ..., w_{i-1})$$

La ecuación anterior presenta una oportunidad de aproximar $P(w_1, w_2, ..., w_m)$ por limitación del contexto:

$$P(w_1, w_2, ..., w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \text{, para cualquier } n \geq 2.$$

Si asumimos el lenguaje como ergódico, es decir, que tiene la propiedad de que la probabilidad de cualquier estado puede ser estimado desde un historial lo suficientemente grande independientemente de las condiciones iniciales. Si *n* es lo suficientemente grande, la ecuación anterior es exacta. Debido a la escasez de datos y a aspectos prácticos de almacenamiento el conjunto de valores usados para *n* es de *1* a *4* inclusive. Los modelos que usan un continuo pero limitado contexto son conocidos como modelos de lenguaje de n-gramas, y la componente de contexto condicional de la probabilidad de la ecuación anterior, $w_{i-n+1}, ..., w_{i-1}$, es conocido como *historial* [S.Young *et al.,*The HTK Book, 2006].

La estimación de las probabilidades en los modelos de n-gramas está generalmente basada en la máxima probabilidad que se puede expresar en términos de ocurrencia en el texto de entrenamiento como:

$$P(w_i \mid w_{i-n+1}, ..., w_{i-1}) = \frac{C(w_{i-n+1}, ..., w_i)}{C(w_{i-n+1}, ..., w_{i-1})}$$

, donde C(.) es el número de veces que aparece una secuencia de palabras dada en el texto de entrenamiento.

La elección de *n* tiene un efecto significativo sobre el número de parámetros potenciales que el modelo pede tener, el cual esta limitado por $|V|^n$, donde V es el diccionario (conjunto de palabras que hay en el LM). Un modelo de 4-gramas con un tamaño de diccionario de 65.000 palabras, puede tener $|65.000|^4 \approx 1.8 \cdot 10^{19}$ parámetros potenciales. Pero en la práctica, sólo un pequeño subconjunto de combinaciones de posibles parámetros representa probablemente las secuencias de palabras. También el requerimiento de

almacenamiento es mucho menor que el máximo teórico (del orden de 10^{11} menos). Incluso dada una significante reducción de la cobertura del idioma y una gran cantidad de datos de entrenamiento, hay todavía posibles secuencias de palabras que no serán encontradas en el texto de entrenamiento o no serán encontradas un numero de veces estadísticamente significativo.

No sólo se debe tener en cuenta el espacio de almacenamiento, sino también hay que ser capaz de conseguir un razonable grado de confianza en las estimaciones obtenidas. Con el incremento de la cantidad de texto de entrenamiento, conseguimos un mayor grado de confianza en las estimaciones del modelo, sin embargo, también es necesario más espacio de almacenamiento y periodos de análisis más largos en la estimación de los parámetros del modelo.

2.6. Reconocimiento basado en N-Best y lattices.

El uso de *lattices* y reconocimiento N-Best es una línea de investigación, utilizada tanto para reconocimiento de habla, como para reconocimiento de idioma o de locutor y *Word-Spotting*. La idea es utilizar modelos acústicos y modelos de idioma para producir una lista con las *n* secuencias de palabras más probables en un tiempo razonable. Entonces, estas *n* hipótesis son re-estimadas usando modelos más precisos para obtener la secuencia de palabras más probable. La idea con las *n* mejores hipótesis (*n-best*) puede ser extendida para crear una representación más compacta de las mismas. N-Best o *lattices* son muy utilizados para sistemas de reconocimiento de habla continuos para varios y amplios vocabularios.

1. I will tell you would I think in my office
2. I will tell you what I think in my office
3. I will tell you when I think in my office
4. I would sell you would I think in my office
5. I would sell you what I think in my office
6. I would sell you when I think in my office
7. I will tell you would I think in my office
8. I will tell you why I think in my office
9. I will tell you what I think on my office
10. I Wilson you I think on my office

Tabla 1- Ejemplo de una lista 10-best para una frase del ámbito de los negocios [Huang *et al.*, 2001; p. 665]

En la *Tabla 1*, se muestra un ejemplo de una lista n-best (10-best) generada para una frase del ámbito de los negocios. Nótese que algunas de las distintas secuencias de palabras difieren únicamente en una palabra. Esto es razonable, ya que similares secuencias de palabras, tendrán similares puntuaciones (probabilidades). En general, el número de hipótesis n-best puede crecer exponencialmente con la longitud de la secuencia. Por eso, los *lattices* de palabras son introducidos, reemplazando las listas n-best por una representación mas compacta de las hipótesis alternativas. Los *lattices* representan los distintos caminos posibles, es decir, las distintas e hipotéticas secuencias de palabras, y sus probabilidades asociadas (*Figura 19*).

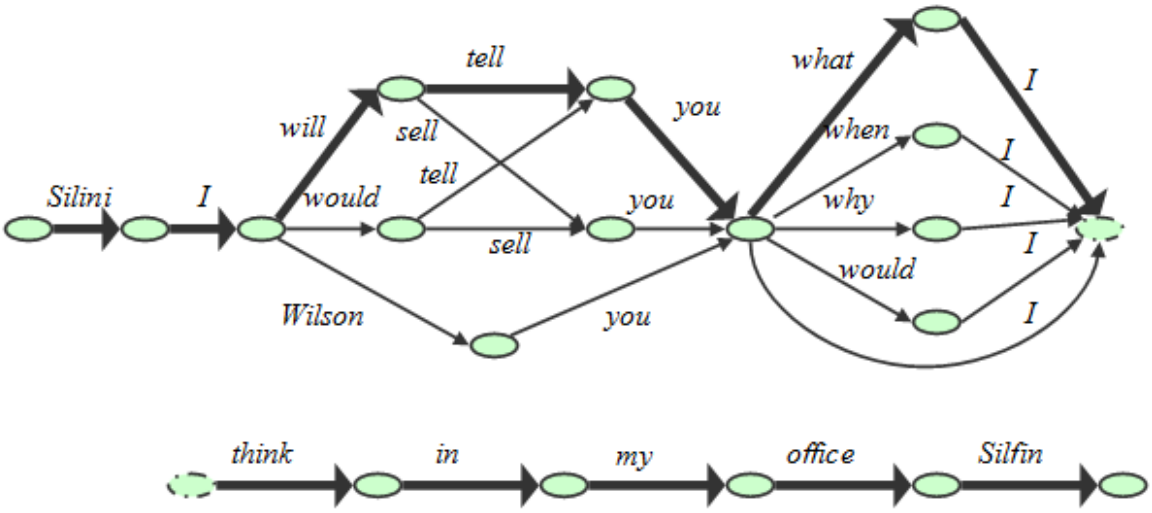

Figura 19- Representación gráfica del lattice de palabras, para la frase del ámbito de negocios.

3. Diseño y desarrollo.

3.1. Medios disponibles.

3.1.1. Bases de datos.

Para la realización de un reconocedor fonético para un idioma concreto se requiere, para la fase de entrenamiento y de evaluación, una colección de archivos de audio del idioma y de ficheros con sus respectivas transcripciones, es decir, con los contenidos de los ficheros de audio. Estas transcripciones pueden encontrarse a nivel de fonemas o nivel de palabras. En este último caso, además se requiere un transcriptor, que permita obtener las transcripciones a nivel de fonemas a partir de las transcripciones a nivel de palabras.

Para el desarrollo de este proyecto se ha hecho uso de la base de datos *SpeechDat-English*.

3.1.2. SpeechDat.

SpeechDat es una colección de bases de datos de habla telefónica de 8 idiomas: danés, inglés, francés, francés de Suiza, alemán, italiano, portugués y español. Con el fin de proporcionar una fuente común para el desarrollo general de tele-servicios de reconocimiento de habla, la especificación es idéntica, o muy parecida, en la mayoría de aspectos para cada una de las bases de datos.

Estas bases de datos han sido creadas para el entrenamiento y evaluación de sistemas de reconocimiento de habla independiente de locutor. Por ello incorpora archivos de audio de 5000 locutores diferentes, logrando una cobertura representativa de los dialectos regionales importantes.

El género, la edad, la educación y estatus socio-económico también contribuyen a las variaciones que ocurren en el habla humano. Estas bases de datos presentan un buen equilibrio en cuanto al género de los locutores, y un amplio y equilibrado rango de edades como se muestra en la siguiente tabla.

Age	Minimun %
<16	optional
16-30	20
31-45	20
46-60	20
>60	optional

Tabla 2- Distribución de las edades de los locutores.

El diseño de *SpeechDat* ha sido motivado por la iniciativa de la base de datos *COCOSDA Polyphone*. Las propiedades técnicas del conjunto de datos de *Polyphone* son:

- 20-40 palabras por locutor (leídas y espontáneas),
- 5000 locutores,

- material de habla telefónica recopilada de forma digital directamente desde la red del teléfono.

El vocabulario de *Polyphone* contiene:

- Palabras orientadas a aplicación.
- Secuencias de números.
- Palabras y nombres deletreados.
- Fechas, tiempos y precios.
- Frases que proporcionan cobertura fonética del idioma.

A cada archivo de audio le corresponde uno de datos, que proporciona diversa información (véase la *Tabla 3*) sobre la transcripción, el locutor, el entorno de grabación, la codificación,… siendo la de mayor importancia la transcripción ortográfica.

LHD: SAM, 6.0 DBN: SpeechDat_East_Russian_Fixed_Network VOL: FIXED3RU_04 SES: 0800 DIR: \FIXED3RU\BLOCK08\SES0800 CMT: ********** File information ********** SRC: A30800A1.RUA CCD: A1 CRP: BEG: 0 END: 45595 ASS: OK REP: AudiTech Ltd, St.Petersburg, Russia RED: 03/October/1999 RET: 14:06:16 12 CMT: ********** Speech data coding ********** SAM: 8000 SNB: 1 SBF: SSB: 8 QNT: A-LAW CMT: ********** Speaker information ********** SCD: 064500 SEX: F AGE: 47 ACC: TVER CMT: ********** Recording condition ********** REG: MIDDLE RUSSIA ENV: HOME NET: FIXED PHM: ROTARY LBD: CMT: ********** BODY ********** LBR: 0,45595,,,,повторить LBO: 0,22798,45595,[sta] повторить [spk] ELF:	mnem. comments LHD: format name + version ELF: end of label file CMT: comment row DBN: database name VOL: database volume ID SES: session number DIR: signal file directory SRC: signal file name CCD: corpus code CRP: corpus repetition BEG: labelled sequence start position END: labelled sequence end position ASS: assessment code REP: recording place: place, city, country RED: recording date RET: recording time (:SS = :00) SAM: sampling frequency SNB: number of (8-bit) bytes per sample SBF: sample byte order (meaningless with single byte samples, .SNB: 1.) SSB: number of significant bits per sample QNT: quantization SCD: speaker code SEX: speaker sex AGE: speaker age ACC: speaker accent REG: calling region ENV: calling environment NET: telephone network PHM: telephone hand set model LBD: label body keyword LBR: labelling during recording: begin, end, gain, min, max, orthographic text prompt LBO: orthographic labelling: begin, centre, end, orthographic transcription text

Tabla 3- Formato del archivo de datos asociado a un archivo de audio [*SpeechDat English*].

Dicha transcripción ortográfica es a nivel de palabras, por lo que, es necesario de alguna manera obtener la transcripción a nivel fonético a partir de ésta, como ya hemos comentado. Para dicha labor, se ha implementado un ***transcriptor fonético de palabras*** descrito con detalle en la *Sección 3.2.1*.

3.1.2.1. Preparación de los datos.

Como ya hemos comentado al principio de la Sección, para la implementación de un reconocedor fonético se necesita las transcripciones de cada uno de los dicheros de audio utilizados en la etapa de entrenamiento y evaluación. Pero para la obtención de buenos resultados, también es imprescindible que las transcripciones, obtenidas de los ficheros de datos descritos en la *Tabla 3*, sean correctas. Por este motivo se realizó un estudio en profundidad de los archivos de audio, de datos y las transcripciones que contienen, obteniendo como resultado las siguientes observaciones *"desfavorables"* para el desarrollo de este proyecto:

- Mala pronunciación de palabras, que no obstante son inteligibles, como consecuencia de los dialectos (por ejemplo,*"cah"* por *"car"*)o de las variaciones del idioma (por ejemplo, *"wanna"* por *"want"*).
- Partes de archivo de audio ininteligibles.
- Fragmentos de palabras que, por ejemplo, el locutor no completa.
- Eventos acústicos no pertenecientes al habla, tales como:
  - Pausas de relleno, que representan eventos acústicos similares acústicamente y fonéticamente al habla (por ejemplo, *[ah]* o *[mm]*).
  - Todos los tipos de ruidos hechos por el locutor (por ejemplo, suspiros o respiración fuerte).
  - Todos los tipos de ruidos no realizados por el locutor (por ejemplo, pasos u otras voces).
- Señales de habla truncadas debido a errores de grabación.

Todos estos hechos observados son representados por caracteres especiales en el campo de transcripción de los archivos de datos, que imposibilitan la automatización del tratamiento de estos ficheros en la obtención de las transcripciones.

Por otra parte, debido al gran número de archivos de la base de datos a utilizar (en torno a los 39.000), se descartó el tratamiento individual de cada uno de los archivos de datos. Finalmente, tras analizar las consecuencias, se optó por la supresión de los archivos que afectados por alguno de las anteriores efectos, salvo aquellos que contuviesen unos determinados tipos de ruido (mostrados en la *Tabla 4*). Se decidió, por tanto, modelar dichos ruidos como fonemas especiales, ya que eran los más frecuentes, consiguiendo así conservar un mayor número de archivos para la implementación del proyecto. Tras estas decisiones se posibilitó la automatización del tratamiento de los archivos de datos en la obtención de las transcripciones de cada uno de los archivos de audio, y manejando finalmente una gran cantidad de archivos (27.988, es decir, el 72% de archivos de la base de datos inicial).

Tipo de ruido	*[ah]*	*tos*	*[er]*	*risas*	*chasqueo de labios*	*[mm]*	*[oh]*	*suspiro*	*[uh]*	*[um]*
Frecuencia	10	141	166	19	13	13	14	44	14	59

Tabla 4- Tipos de ruidos modelados y su frecuencia.

Normalmente, las bases de datos están divididas en dos directorios, uno que contiene los archivos a utilizar en la fase de entrenamiento y otra con los archivos a utilizar en la fase de evaluación. Sin embargo, *SpeechDat-English* no tiene tal división, por tanto se ha tomado un 80% de los archivos de audio para el entrenamiento y el 20% restante para la evaluación.

Por último, destacar que los archivos de audio de esta base de datos vienen en formato *raw* comprimido y codificado en formato ley A, muestreado a 8 KHz y mono. Por ello, antes de empezar a usarlos, es necesario transformarlos a archivos *'.wav'*, en formato PCM y muestreados a 8 KHz para que pedan ser procesados por el parametrizador.

3.1.3. Software.

El sistema operativo empleado en la realización de este proyecto ha sido Debian y la principal herramienta de *software* empleadas son HTK.

El *Hidden Markov Model Toolkit* (HTK) es un *toolkit* portable usado para la construcción y manipulación de los modelos de Markov. HTK fue usado en principio en aplicaciones de reconocimiento de voz, aunque se ha encontrado otras muchas aplicaciones como la síntesis de voz o secuencias de ADN.

HTK consiste en un conjunto de librerías y herramientas desarrolladas en C. Las sofisticadas herramientas desarrolladas facilitan el análisis de la voz, el entrenamiento de los HMM, la evaluación y extracción de resultados. El *software* soporta la creación de HMM con distribuciones continuas de mezclas de Gaussianas o por medio de distribuciones discretas pudiendo crear de esta forma complejos sistemas de HMM.

HTK fue desarrollado originalmente en el laboratorio de la inteligencia de máquinas (conocido antes como el grupo "*the Speech Vision and Robotics*") del departamento de ingeniería de la *Universidad de Cambridge* (CUED) donde se ha utilizado para construir grandes sistemas de reconocimiento de habla. En 1993 *Entropic Research Laboratory Inc*. adquirió los derechos de vender HTK y el desarrollo de HTK fue transferido completamente a *Entropic* en 1995 en que el laboratorio de investigación de *Entropic Cambridge Ltd* fue establecido. HTK fue vendido por *Entropic* hasta que en 1999 Microsoft compró *Entropic*. Microsoft ahora ha licenciado HTK de nuevo a CUED y está proporcionando la ayuda de modo que CUED pueda redistribuir HTK

Otro programa utilizado para el desarrollo de este proyecto es *Sphinx*. El grupo *Sphinx* está desarrollado por la *Universidad Carnegie-Mellon*, la *Defense Advanced Research* Projects Agency (DARPA) financia extensamente el proyecto para estimular la creación de herramientas de discurso y el uso de las mismas, en el reconocimiento de voz, así como en áreas relacionadas incluyendo sistemas de diálogo y síntesis de discurso.

Sphinx ha sido apoyado durante muchos años por la financiación del DARPA y los motores del reconocimiento que se lanzan son los que el grupo utilizó para varios de los proyectos de DARPA y sus evaluaciones respectivas.

La ayuda reciente para el proyecto también incluye *Telefónica I + D*, *Sun Microsystems*, y los laboratorios de investigación eléctricos de *Mitsubishi.*

Los términos en que se licencian para los motores y las herramientas de *Sphinx* se derivan del DEB, y se basan, particularmente, sobre la licencia para el *web server* de Apache. No hay restricción contra uso o la redistribución comercial.

Los paquetes que el grupo *Sphinx* de CMU está lanzando son un sistema razonablemente maduro. Los componentes proporcionan un nivel básico de la tecnología a cualquier persona interesada en crear reconocedores sin el coste de inversión inicial; los mismos componentes están abiertos a la revisión para todos los investigadores en el campo, y se utilizan también para la investigación lingüística. [*Carnegie Mellon University* SPHINX]

Como herramientas de trabajo se han utilizado en el desarrollo del proyecto distintos lenguajes de programación como Perl, bash y C.

3.1.4. Hardware.

El *hardware* empleado en el desarrollo de este proyecto ha sido un ordenador con procesador Intel Pentium IV. Además, he tenido a mi disposición una red interna formada todos los ordenadores del grupo de trabajo, tanto los de uso personal como los de pruebas. Todos estos medios fueron suministrados por el grupo ATVS de la Universidad Autónoma de Madrid (UAM).

3.2. Diseño.

3.2.1. Transcriptor fonético de palabras para inglés americano.

Debido al requerimiento de la transcripción fonética de los archivos de audio de la base de datos en las fases de entrenamiento y evaluación, hemos creado un *transcriptor fonético de palabras para inglés americano*.

El transcriptor fonético de palabras implementado transcribe fonéticamente las palabras contenidas en el archivo de entrada, obteniendo un archivo de salida con las transcripciones fonéticas de las palabras de entrada.

El conjunto de fonemas usado por el transcriptor es el siguiente:

aa	*aw*	*d*	*ey*	*ih*	*l*	*ow*	*s*	*uh*	*y*
ae	*ay*	*dh*	*f*	*iy*	*m*	*oy*	*sh*	*uw*	*z*
ah	*b*	*eh*	*g*	*jh*	*n*	*p*	*t*	*v*	*zh*
ao	*ch*	*er*	*hh*	*k*	*ng*	*r*	*th*	*w*	

Tabla 5- Conjunto de fonemas elegidos.

Este conjunto de fonemas permite representar los caracteres *IPA (International Phonemes Association)* como se muestra en la T*abla 6*. Los caracteres IPA son la representación estandarizada de los sonidos del habla.

I	iy	ɛ	eh	ɔ	ao	ɪ	ih r	ɛɹ	eh r	P	p	D	d	M	m
U	uw	ɝ	er	ɔɹ	ao r	ʊɹ	uh r	ɔɪ	oy	B	b	ʤ	jh	N	n
ɪ	ih	ə	ah	ɑɹ	aa r	Eɪ	ey	aɪ	ay	T	t	K	k	Ŋ	ng
ʊ	uh	Æ	ae	ɑ	aa	Oʊ	ow	aʊ	aw	ʧ	ch	G	g	F	f
Z	z	ʃ	sh	ʒ	zh	h	hh	ɹ	r	J	y	L	l	W	w
V	v	Θ	th	Ð	dh	S	s	Representación de los caracteres IPA							

Tabla 6- Representación de los caracteres IPA con el conjunto de fonemas del diseño.

Para la implementación de este transcriptor nos hemos servido del diccionario de pronunciación de CMU, conocido como *cmudict*. Este diccionario de dominio público fue creado por la *Carnegie Mellon University*. La versión utilizada es la 0.6 que contiene 127.069 entradas y hace uso del mismo conjunto de fonemas que el escogido para el desarrollo de proyecto. Sus transcripciones también contienen marcas de acentuación representadas con 0, 1 o 2.

Haciendo uso del *cmudict* podemos obtener la transcripción fonética de un gran número de palabras (de 127.069 palabras). Para dotar de total cobertura al transcriptor, hemos utilizado un *software* creado por *NIST (National Institute of Standars and technology)*, llamado *'addttp4'*. Dicho *software*, escrito en ANSI C, transcribe texto a fonemas mediante un conjunto de reglas *TTP (Text-To-Phone)* y parámetros de silabificación.

El *software* de NIST utiliza un conjunto de fonemas que difiere al conjunto de fonemas elegido para el desarrollo del proyecto, aunque es parecido. Para hacer compatible el uso conjunto del diccionario de CMU y del *software* de NIST se ha implementado un modulo que mapea los fonemas de las transcripciones de salida de *'addttp4'* que no pertenecen a nuestro conjunto de fonemas. Los fonemas que maneja el *software* de NIST que no pertenecen al conjunto de fonemas del diseño son dos: *nx* que se corresponde directamente con el fonema *ng* de nuestro diseño y *ax* que no tiene ninguna correspondencia directa con ninguno de los fonemas del conjunto de diseño. Pero tras un estudio realizado con una gran cantidad de transcripciones, se ha decidido mapear el fonema *ax* con el fonema *ah*.

Por otra parte, cuando el *software 'addttp4'* no consigue separar una palabra en sílabas (haciendo uso de sus parámetros de silabificación) la transcribe como si fuese un acrónimo, es decir, la deletrea. Y se observó que esto ocurre con demasiada frecuencia en palabras que no son acrónimos, por lo que se decidió modificar el código del programa para solventar dicho problema.

Un último tema que se trató fue respecto a las marcas de acentuación (*stress*). Se añadió la opción de que el transcriptor maneje o no marcas de acentuación. El reconocedor de habla implementado no maneja dichas marcas de acentuación, pero puede resultar interesante para otros proyectos mantener la posibilidad de que el transcriptor fonético de palabras maneje marcas de acentuación.

De este modo, el funcionamiento del transcriptor fonético implementado se puede resumir en los siguientes pasos:

1. Dado un archivo con la lista de palabras que se desea transcribir, el transcriptor lee cada una de las palabras de entrada.

2. Busca la palabra de entrada en el diccionario *cmudict*.

3. Si encuentra la palabra en el diccionario, guarda la transcripción fonética que la corresponde en el archivo de salida, y vuelve al *paso 2* con la siguiente palabra de la lista de entrada.

4. Si no encuentra la palabra en el diccionario, entonces ejecuta el *software* '*addttp4*', que obtiene su transcripción. A continuación realiza sobre la transcripción el mapeo de los fonemas para adaptarlo al nuestro conjunto de fonemas. Después guarda la transcripción fonética correspondiente en el archivo de salida y vuelve al *paso 2* con la siguiente palabra de la lista de entrada.

En la que se muestra a continuación (*Figura 20*) se representa el funcionamiento del transcriptor mediante un diagrama de bloques.

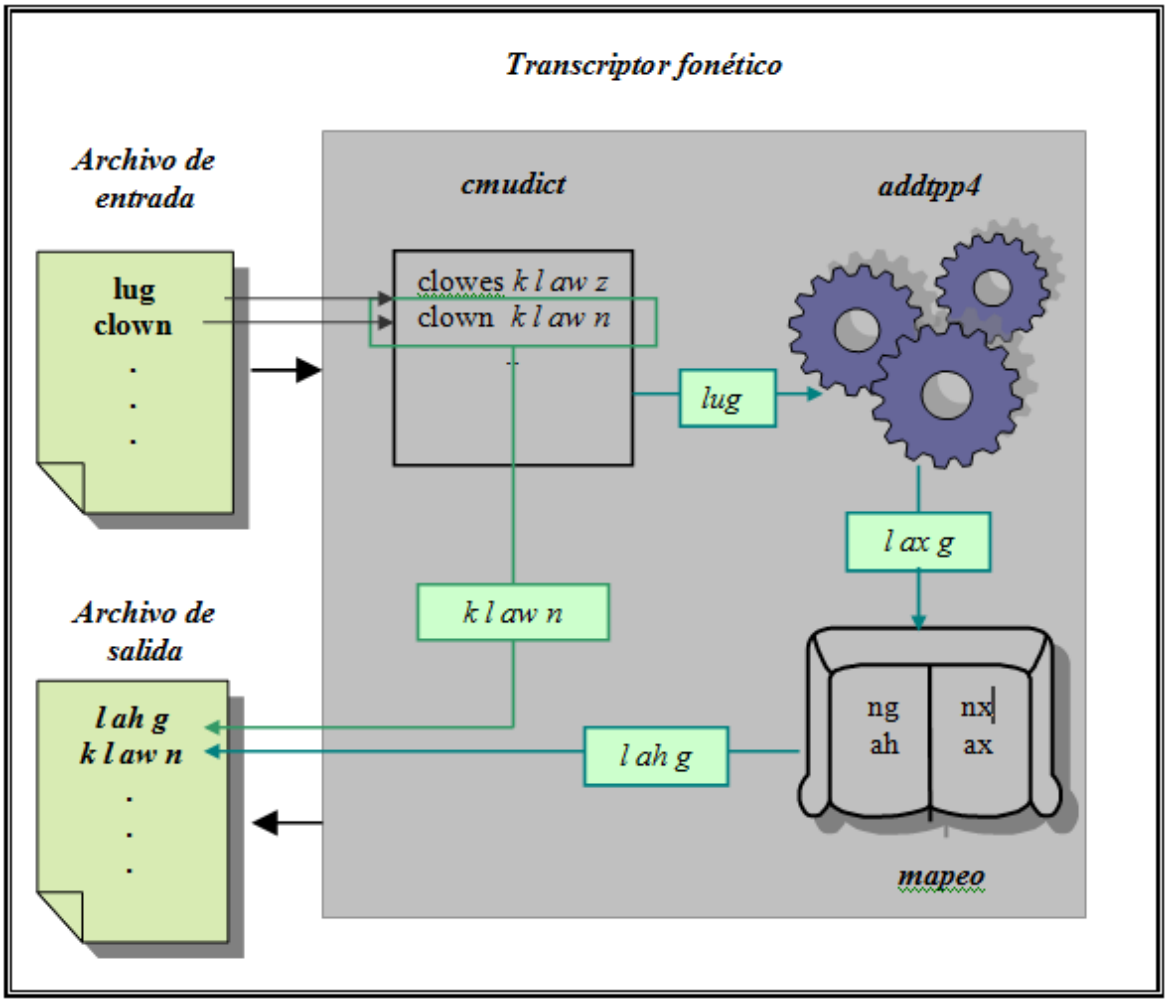

Figura 20- Diagrama de bloques que representa el funcionamiento del transcriptor fonético de palabras.

3.2.2. Parametrización de las señales de audio.

En primer lugar se parametrizaron todos los archivos de audio de la base de datos. Para ello, previamente se tuvo que cambiar el formato de los archivos de audio – archivos 'raw' comprimidos y codificados según la ley A (archivos de habla telefónica) a archivos 'wav ' con una codificación PCM de 16 bits.
Una vez cambiado el formato de todos los archivos de audio, se parametrizaron con *Sphinx*. Dicha parametrización se basa en la aplicación de un filtro de pre-énfasis a cada una de las señales de habla y su posterior análisis utilizando ventanas Hamming de 25 ms con 10 ms de desplazamiento entre ventanas, obteniendo 13 coeficientes MFCC por cada ventana. Como resultado de la parametrización es un fichero por cada archivo de audio con sus coeficientes correspondientes.

Para posibilitar el uso de estos archivos con HTK se añadieron los coeficientes MFCCs-Delta (13 coeficientes) y los MFCCs-Delta-Delta (13 coeficientes). Expandiendo el número de coeficientes por ventana a 39.

3.2.3. Creación de los modelos acústicos independientes del contexto.

Para nuestro reconocedor de habla desarrollamos modelos acústicos independientes del contexto mediante los HMMs. La topología de estos modelos se muestra en la *Figura 21*, son 3 estados de izquierda a derecha sin saltos.

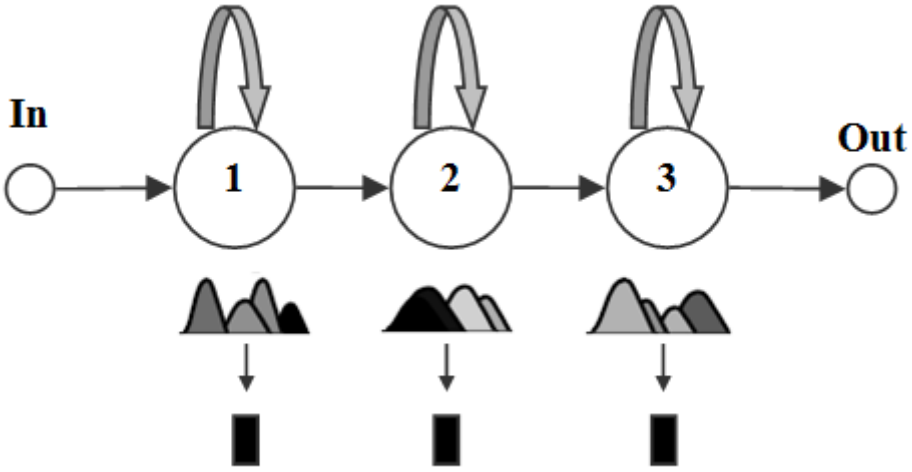

Figura 21- Topología del modelo acústico de los fonemas.

Para cada fonema de nuestro conjunto definido (incluidos los ruidos que se decidieron modelar) se tendrá un modelo con el formato mostrado en la *Figura 21*, donde los estados están representados por una Gaussiana y cada vector tiene una longitud de 39. Este número, 39, es el conjunto de los 13 coeficientes MFCC, los 13 coeficientes MFCC-Delta y los 13 coeficientes MFCC-Delta-Delta. Después en el entrenamiento, como se explica en la siguiente Sección, se añadirán más Gaussianas, de manera que cada estado estará representado por una mezcla de Gassianas, con el fin de mejorar la estimación de los modelos acústicos. La matriz de *5x5* indica las probabilidades de transición entre estados.

```
<BeginHMM>
 <NumStates> 5
 <State> 2
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <State> 3
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <State> 4
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <TransP> 5
  0.0 1.0 0.0 0.0 0.0
  0.0 0.6 0.4 0.0 0.0
  0.0 0.0 0.6 0.4 0.0
  0.0 0.0 0.0 0.7 0.3
  0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Figura 22- Formato HTK del modelo acústico de los fonemas.

Una de las herramientas de HTK escanea el conjunto de archivos de la base de datos y computa la media y la varianza global y, dado un HMM, pone todas las Gaussianas con la misma media y varianza. El modelo obtenido será el modelo de partida para cada uno de los fonemas de nuestro conjunto.

A parte de estos modelos, utilizamos tres modelos de silencio: silencio inicial, silencio final, y pausa corta. La topología de los dos primeros es idéntica a la del modelo de los fonemas, salvo por una transición ente los estados emisores 1 y 3, permitiendo mayor flexibilidad en el proceso de modelado del silencio. El modelo de pausa corta tiene sólo un estado emisor (*Figura 23*), ya que modela mejor su típica duración corta. Este modelo, que se creará mas adelante, se conoce como *tee-model y* tiene una directa transición directa entre su estado de entrada y salida.

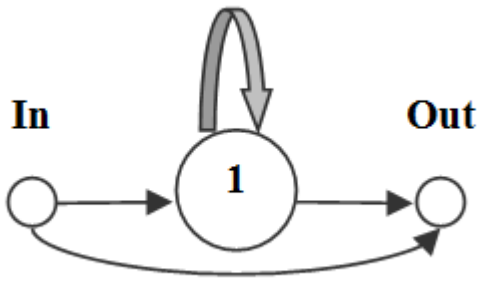

Figura 23- Topología del modelo acústico para la pausa corta.

Partiendo de los modelos acústicos creados para los fonemas (incluidos los ruidos) y para el silencio final e inicial, se realizaron 4 re-estimaciones con los datos de entrenamiento, mediante las herramientas de HTK. El siguiente paso fue añadir transiciones extras desde el estado 1 al 3 y desde el 3 al 1 en los modelos del silencio final e inicial. Con ello se pretende hacer un modelo más robusto permitiendo estados individuales para absorber los distintos ruidos impulsivos de los datos de entrenamiento. El salto hacia atrás permite que puedan ocurrir silencios sucesivos sin tener que ser asignados a otras palabras.

En este punto, se creó el modelo para la pausa corta, con los parámetros del estado central del modelo acústico de uno de los silencios. Y por último se ató su estado emisor con el estado central de los silencios inicial y final. Dicha topología se muestra en la *Figura 24*.

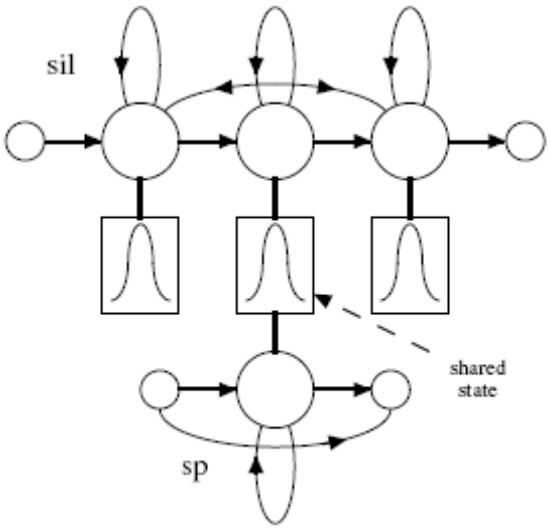

Figura 24- Topología del modelo de silencio [HTK Book, p.33].

A continuación se realizaron 4 re-estimaciones adicionales con los datos de entrenamiento.

3.2.4. Creación de los modelos acústicos dependientes del contexto - Tri-fonemas.

Un modelo de un tri-fonema es un modelo fonético que tiene en cuenta el fonema vecino de la izquierda y de la derecha. Si dos fonemas tienen la misma identidad pero diferente contexto de la izquierda y/o de la derecha, entonces, son considerados tri-fonemas distintos. Los modelos basados en tri-fonemas son potentes porque capturan los efectos más importantes de la coarticulación. Y son generalmente más consistentes que los modelos fonéticos independientes del contexto.

Dado el conjunto de HMMs de fonemas se crearon HMMs de tri-fonemas dependientes de contexto, haciendo uso de las herramientas de HTK y se realizaron 4 re-estimaciones con los datos de entrenamiento.

Para crear estos modelos, HTK opera de la siguiente manera: para cada modelo de la forma */a/-/b/+/c/* busca el modelo de */b/* y lo copia, mientras que la matriz de transición pasa a ser un subcomponente del HMM, y todas las matrices de transición individuales de todos los tri-fonemas cuyo fonema central es */b/,* es remplazada por una macro, que es compartida por todos los modelos de dichos tri-fonemas (*Figura 25*). La razón de estas macros es que los parámetros de transición no varían significadamente con el contexto acústico.

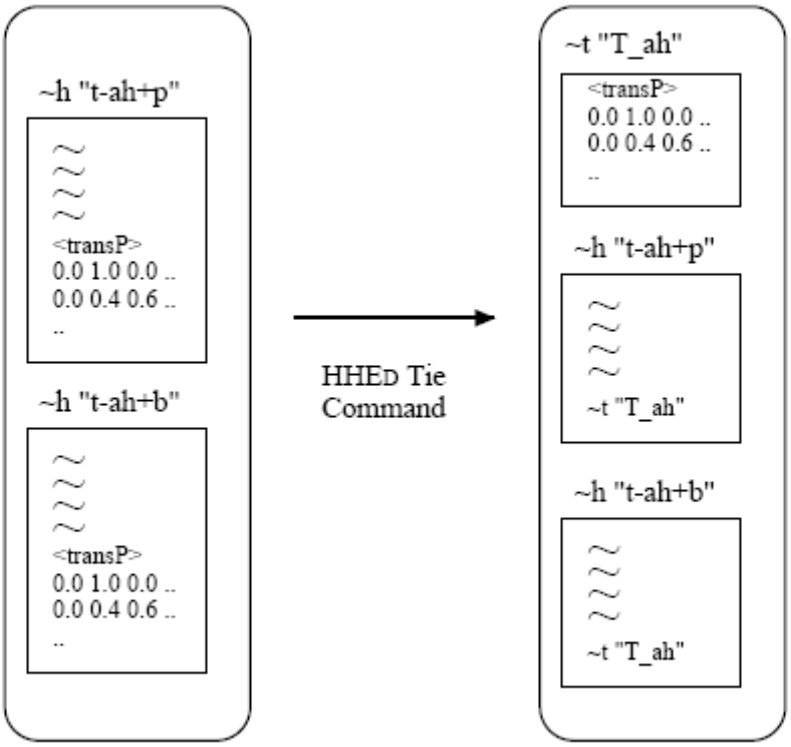

Figura 25- Matrices de transición compartidas [HTK Book, p.36].

3.2.5. Construcción de la gramática.

La gramática es un conjunto de reglas que limita el número de combinaciones de fonemas, tri-fonemas o palabras, permitidas, utilizada en el reconocimiento, con el fin de mejorar la tasa de reconocimiento a través de la eliminación de ambigüedades, además de, aumentar la rapidez y precisión del proceso. La gramática a construir depende del tipo de reconocimiento (fonético, a nivel de palabras, etc) a realizar y de la mejora que se quiera introducir en el sistema. Por tanto, para los diversos experimentos llevados a cabo en este proyecto, se crearon dos gramáticas distintas mediante las herramientas de HTK: una a nivel de fonemas y otra a nivel de palabras, para el reconocimiento fonético y de palabras respectivamente.
También se estudió la creación de otras dos gramáticas (véase la *Sección 4.2.4*), una a nivel de tri-fonemas para reconocimiento a nivel de tri-fonemas y otra basada en el modelo de idioma.

3.2.6. Sistema de reconocimiento fonético.

Para la implementación final del sistema de reconocimiento fonético se llevaron a cabo las siguientes operaciones.

3.2.6.1. Obtención de las transcripciones fonéticas.

Tanto para la fase de entrenamiento de los modelos fonéticos creados, como para la fase de evaluación del sistema, se requieren las transcripciones fonéticas de los archivos de audio de la base de datos. Dichas transcripciones, se crean a partir de las transcripciones a nivel de palabras contenidas en los archivos de datos y de un ***diccionario fonético*** que contenga todas las palabras manejadas en la base de datos y sus transcripciones fonéticas.

Para la creación del diccionario, se hizo uso del transcriptor fonético de palabras descrito en la *Sección 3.2.1*.: se recopilaron todas las palabras manejadas en la base de datos y mediante el transcriptor, se obtuvieron sus transcripciones fonéticas.

A continuación, se creó un archivo con la transcripción a nivel de palabras de todos los archivos de audio de la base de datos en el formato MLF (*Macro Label File*) de HTK. A partir de este archivo y haciendo uso del diccionario fonético y de las herramientas de HTK, se obtuvo el archivo en formato HTK con todas las transcripciones fonéticas de los archivos de la base de datos.

3.2.6.2. Entrenamiento de los modelos acústicos fonéticos.

Una vez obtenidas las transcripciones fonéticas, se pasó a la fase final de entrenamiento de los modelos fonéticos creados, haciendo uso de las herramientas de HTK. Esta fase consiste en añadir Gaussianas a todos los estados de todos los modelos. Por cada Gaussiana añadida se realizaron otras 4 re-estimaciones con los datos de entrenamiento. El número final de Gaussianas por estado que se alcanzó fue de 39. En número de Gaussianas se determinó teniendo en cuenta la relación mejora-tiempo de computación.

3.2.6.3. Construcción de la gramática basada en fonemas.

Para la realización de las pruebas de reconocimiento de este sistema se construyó una gramáticas a nivel de fonemas (véase la *Sección 4.2.1*). En su construcción no se tuvo en cuenta el conocimiento léxico-semántico de la lengua, es decir, todos los fonemas tienen la misma probabilidad de aparición tras un fonema dado.

3.2.7. Sistema de reconocimiento de palabras basado en modelos fonéticos.

Para la implementación de este sistema de reconocimiento a nivel de palabras se hizo uso de los modelos fonéticos entrenados para el sistema de reconocimiento fonético, sin embargo, para las pruebas de reconocimiento se creó una gramática a nivel de palabras.

3.2.7.1. Gramática basada en palabras.

Para la realización de las pruebas de reconocimiento de este sistema se construyó una gramáticas a nivel de fonemas (véase la *Sección 4.2.1*). En su construcción no se tuvo en cuenta el conocimiento léxico-semántico de la lengua, es decir, todas las palabras tienen la misma probabilidad de aparición tras una palabra dada.

3.2.8. Sistema de reconocimiento de palabras basado en modelos de tri-fonemas.

Para la implementación de este sistema, se requirió la obtención de las transcripciones a nivel de tri-fonemas de los archivos de audio de la base de datos necesarias, tanto en el entrenamiento de los modelos de tri-fonemas, como en la evaluación del sistema, para la que se utilizará la gramática basada en palabras descrita en la *Sección 3.2.6.3*.

3.2.8.1. Obtención de las transcripciones a nivel de tri-fonemas.

En primer lugar, se convirtieron las transcripciones fonéticas de los archivos de las bases de datos que se tenían, en transcripciones a nivel de tri-fonemas. De esta manera, si antes se tenía, por ejemplo

/silini/ /th/ /ih/ /s/ /sp/ /m/ /ae/ /n/ /sp/ …

tras dicha conversión se obtuvo

/silini/ /th/+/ih/ /th/-/ih/+/s/ /ih/-/s/ /sp/ /m/+/ae/ /m/-/ae/+/n/ /ae/-/n/ /sp/… .

Para ello, se utilizaron las herramientas proporcionadas en HTK, obteniendo un archivo con todas las transcripciones a nivel de tri- fonemas de la base de datos en el formato MLF de HTK.

3.2.8.2. Entrenamiento de los modelos acústicos basados en tri-fonemas.

Algunos tri-fonemas sólo ocurren una o dos veces en los datos de entrenamiento, lo que da lugar a estimaciones muy pobre. Para solventar este problema de insuficiencia de datos se recurrió a atar estados entre conjunto de tri-fonemas con el fin de compartir datos y así realizar una estimación de parámetros robusta. Este atado de estados se basa en que hay

algunos fonemas que tienen efectos sobre sus fonemas vecinos. Por ejemplo, /b/ y /p/ son consonantes labiales y tienen un efecto similar sobre la siguiente vocal, al igual que las dos consonantes líquidas /w/ y /r/ (*Figura 26*).

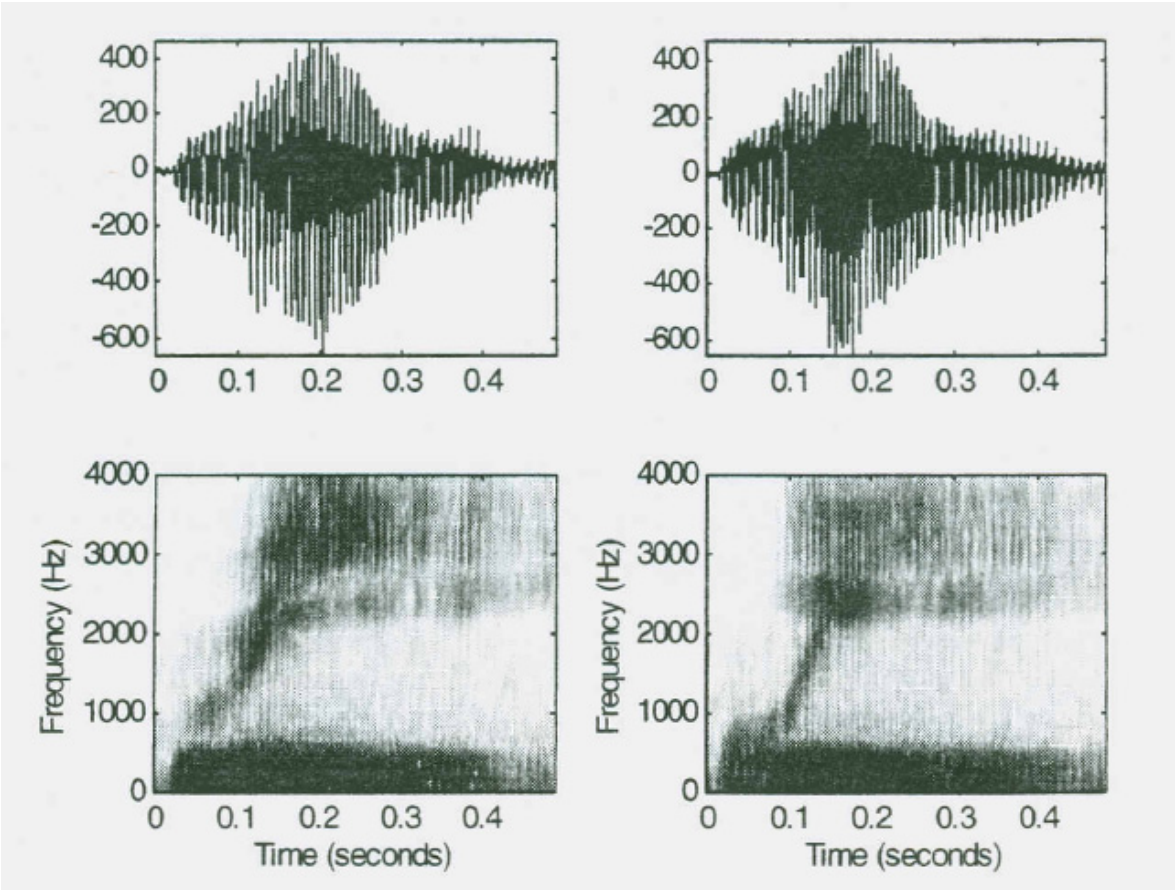

Figura 26- Espectrograma para el fonema /iy/ con dos fonemas vecinos a la izquierda distintos: /r/ y /w/ [Huang *et al.*, 2001; p. 432].

Las características de los fonemas que se tuvieron en cuenta en la atadura de estados fueron las siguientes:

- Si representan vocales o consonantes.
- En cuanto a los fonemas que representan vocales: si la vocal que representa es aguda o grave, o si es cerrada, media o media abierta, tal y como muestra la *Tabla 7*.

Aguda	/ih/, /iy/, /ay/, /oy/, /ey/, /eh/, /er/
Grave	/uh/, /uw/, /ow/, /ao/, /aa/,/ae/, /aw/
Cerrada	/ih/, /iy/, /ey/, /ay/, /oy/, /uh/, /uw/, /ow/
Media	/eh/, /er/, /ah/, /ao/
Media-Abierta	/ah/, /ao/, /ae/, /aa/, /aw/

Tabla 7- Características de los fonemas que representan vocales del habla inglesa.

- En cuanto a los que representan consonantes: se han clasificado según el articulador y el lugar de articulación, como se muestra en la *Tabla 8*, y si son sonoras o sordas (*Tabla 9*).

	Labial	**Dental**	**Palatal**	**Velar**
Oclusiva	/p/, /b/	/t/, /d/		/k/, /g/
Fricativa	/f/	/dh/, /th/, /z/, /s/	/ch/, /sh/, /zh/	/hh/
Nasal	/m/	/n/		/ng/
Líquida		/r/, /l/	/jh/,/y/	
Vibrante	/r/			
Lateral		/l/	/y/	

Tabla 8- Consonantes del habla inglesa según el lugar de articulación (columnas) y la clase de articulador (filas).

Sonora	/b/, /d/, /dh/, /g/, /jh/, /l/, /m/, /n/, /ng/, /r/, /v/, /w/, /y/
Sorda	/ch/, /f/, /hh/, /k/, /p/, /s/, /sh/, /t/, /th/, /z/, /zh/

Tabla 9- Clasificación de las consonantes del habla inglesa.

Para finalizar el entrenamiento de los modelos acústicos basados en tri-fonemas, se añadieron Gaussianas a todos los estados de todos los modelos. Por cada Gaussiana añadida se realizaron otras 4 re-estimaciones con los datos de entrenamiento. El número final de Gaussianas por estado que se alcanzó fue de 39. En número de Gaussianas se determinó teniendo en cuenta la relación mejora-tiempo de computación.

3.2.9. Evaluación de los modelos acústicos.

En la evaluación se utilizan los datos de evaluación de la base de datos (20%) y se compara las transcripciones obtenidas (a nivel de palabras o de fonemas, según el experimento) con las verdaderas transcripciones de los datos, mediante las herramientas de HTK.

La obtención de las transcripciones se obtiene mediante una herramienta de HTK, basado en el *algoritmo de Viterbi* y que dado un conjunto de parámetros de entrada elige la secuencia (de palabras o fonemas) de mayor probabilidad. Y otra herramienta de HTK es la encargada de la compasión y obtención de los resultados. El formato de la presentación de resultados, mostrado en la *Figura 27*, es un modelo de medida estandarizado por NIST.

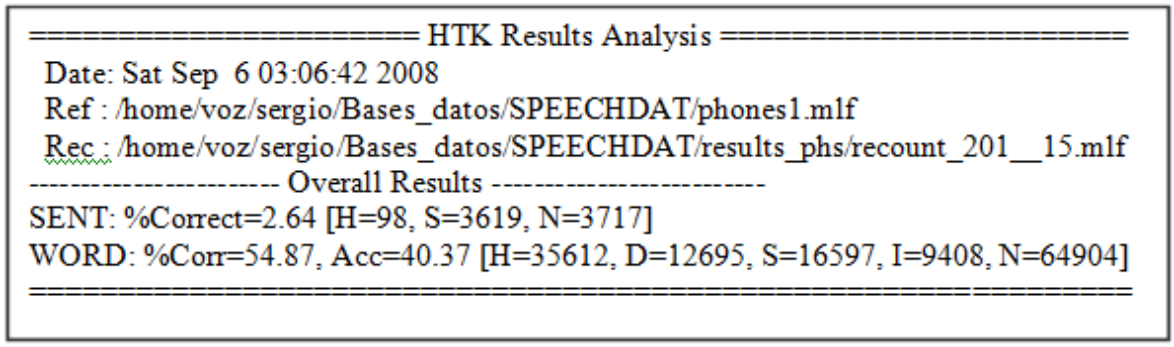

Figura 27- Formato estandarizado por NIST para la representación de resultados de reconocimiento de habla.

Los datos mostrados en la línea cabezada por 'SENT' no son de gran relevancia, ya que indican el porcentaje de frases transcritas que coinciden con las frases originales y suelen ser muy bajos en reconocimiento de habla a nivel de fonemas o de palabras. Los datos representados en la línea encabezada por 'WORD' hacen referencia a medidas a nivel de palabras o fonemas, dependiendo del tipo del reconocimiento. Los valores mostrados son los siguientes:

- **H:** número de fonemas/palabras correctos/as.
- **D:** número de fonemas/palabras borrados/as.
- **S:** número de fonemas/palabras sustituidos/as.
- **I:** número de fonemas/palabras insertados/as.
- **N:** número de fonemas/palabras totales en la transcripción original.
- **%Corr:** porcentaje de fonemas/palabras correctos/as, $\%Corr = \frac{H}{N} x100$
- **%Acc:** precisión del sistema de reconocimiento, $\%Acc = \frac{H-1}{N} x100$

En este proyecto se busca un sistema de reconocimiento con un alto valor de *%Corr*, sin descuidar el valor de *%Acc*, es decir, un alto porcentaje de aciertos que no sea debido a la introducción de un gran número de fonemas espurios. El número de inserciones se ajusta mediante un parámetro de penalización en la herramienta de HTK utilizada para la obtención de las trascripciones.

En ninguna de las evaluaciones realizadas (véase la *Sección 4*) se ha tenido en cuenta la identificación de los silencios, ni de los tipos de silencio, ni de los ruidos modelados.

En la *Figura 28*, se muestra un diagrama de bloques que resume los pasos seguidos para el entrenamiento de los modelos acústicos y evaluación de los distintos sistemas de reconocimiento.

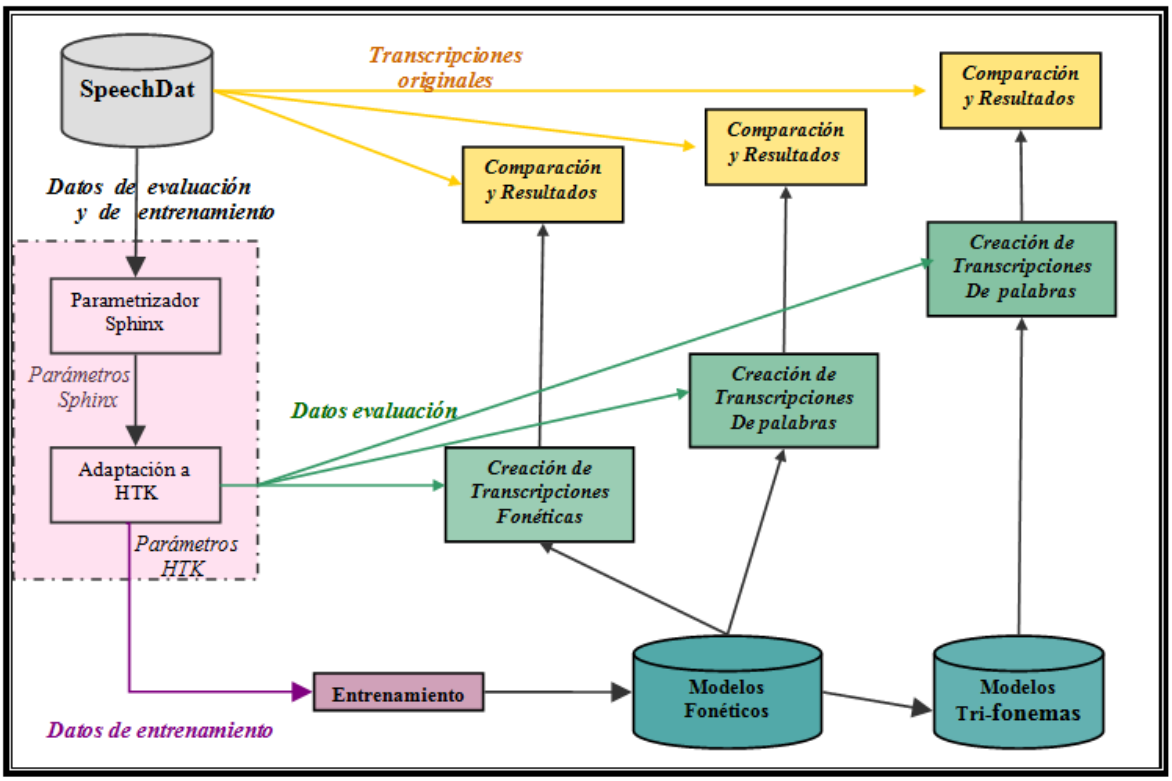

Figura 28- Diagrama de bloques del proceso de entrenamiento y evaluación de los modelos acústicos.

4. Integración, pruebas y resultados

En esta Sección se describen las distintas pruebas realizadas, así como los resultados obtenidos de las mismas, mediante las herramientas proporcionadas por HTK y haciendo uso de los datos de evaluación de la base de datos.

4.1. Reconocimiento fonético.

La primera prueba que se llevo a cabo, fue el reconocimiento fonético utilizando los modelos acústicos obtenidos en la fase de entrenamiento con 39 Gaussianas. Para ello, se hizo uso de:

- una gramática a nivel de fonemas, en la que todos los fonemas tienen la misma probabilidad de aparición tras un fonema dado, como se comentó en la *Sección 3.2.6.3*,
- y de un diccionario que contiene el conjunto de fonemas, incluidos los distintos ruidos y silencios.

Los resultados obtenidos en dicha prueba (*Figura 29)*, muestran un *54.87 %* de fonemas correctos, una alta precisión, *40.37 %,* pero sin embargo se ha obtenido un porcentaje de frases completas correctas muy bajo de *2.64 %*. Esto es razonable, ya que al ser reconocimiento a nivel de fonemas, la probabilidad de que al menos un fonema de toda frase sea erróneo es bastante alta.

```
====================== HTK Results Analysis =======================
  Date: Sat Sep  6 03:06:42 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/phones1.mlf
  Rec : /home/voz/sergio/Bases_datos/SPEECHDAT/results_phs/recount_201__15.mlf
---------------------- Overall Results --------------------------
SENT: %Correct=2.64 [H=98, S=3619, N=3717]
WORD: %Corr=54.87, Acc=40.37 [H=35612, D=12695, S=16597, I=9408, N=64904]
===================================================================
```

Figura 29- Resultados obtenidos en reconocimiento fonético con los modelos fonéticos de 39 Gaussianas obtenidos.

Los mejores resultados publicados de reconocimiento fonético sobre habla leída, microfónica y de *16 KHz,* están en torno a un 75.6 % de precisión fonética. Por lo tanto, los resultados obtenidos son razonables si consideramos que el reconocimiento se lleva a cabo sobre habla telefónica, más o menos espontánea y de *8 KHz*, lo que limita todavía más resultados.

4.2. Reconocimiento a nivel de palabras.

La segunda prueba que se realizó fue el reconocimiento a nivel de palabras con los datos de evaluación de la base de datos. Este reconocimiento se realizó haciendo uso, por un lado, de los modelos fonéticos y por otro, de los modelos basados en tri-fonemas.

4.2.1. Reconocimiento basado en modelos fonéticos.

En estas pruebas de reconocimiento a nivel de palabras, se utilizaron los mismos modelos acústicos que en el reconocimiento fonético anterior, es decir, modelos fonéticos con 39 Gaussianas. Para el desarrollo de dichas pruebas se hizo uso de:

- una gramática a nivel de palabras, en la que todas las palabras tienen la misma probabilidad de aparición tras una palabra dada, como se comentó en la *Sección 3.2.7.1,*
- y de un diccionario que contiene el conjunto de palabras contenidas en la base de datos y sus correspondientes transcripciones fonéticas. Este diccionario se creó, como ya se comento en la Sección 3.2.6.1, haciendo uso del transcriptor fonético implementado en el desarrollo de este proyecto.

```
about       ah b aw t sp
above       ah b ah v sp
absorbed    ah b z ao r b d sp
account     ah k aw n t sp
activate    ae k t ah v ey t sp
activity    ae k t ih v ah t iy sp
actual      ae k ch ah w ah l sp
      .
      .
      .
```

Figura 30- Algunas entradas del diccionario fonético.

Los resultados obtenidos en dicha prueba (*Figura 31)*, muestran un *38.6%* de palabras correctas, una precisión de *25.18%* y un porcentaje de frases completas correctas de *16.60%.*

```
====================== HTK Results Analysis =======================
  Date: Mon Jul 14 00:37:13 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/Ngramas/words_test.mlf
  Rec : /media/datos/BD/LM/ph_recognition/201__45/recout_20_100.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=16.60 [H=929, S=4668, N=5597]
WORD: %Corr=38.36, Acc=25.18 [H=10220, D=3107, S=13318, I=3510, N=26645]
====================================================================
```

Figura 31- Resultados obtenidos en reconocimiento a nivel de palabras con los modelos fonéticos de 39 Gaussianas obtenidos.

Nótese que el porcentaje de palabras correctas, *38.6%*, es menor que el de fonemas correctos de la prueba anterior (véase la *Figura 29*), *54.87 %*, y que la precisión también es menor. Esto es razonable si se analiza de la siguiente manera:

Dado un archivo de audio de los datos de evaluación, en el que se pronuncia *monkey*, cuya transcripción original es *'/m/ /ah/ /ng/ /k/ /iy/'* y supongamos que obtenemos del reconocimiento fonético *'/d/ /ah/ /ng/ /k/ /iy/'*. Entonces en la evaluación del reconocimiento fonético se computarán 4 fonemas correctos y 1 fonema erróneo, mientras que en el reconocimiento a nivel de palabras se computará una palabra errónea.

Nótese también, que a pesar del menor porcentaje de palabras correctas obtenido, el porcentaje de frases completas correctas ha aumentado de un 2.64% (en el reconocimiento fonético) a un *16.60%*, lo cual también es razonable. Supongamos una frase formada, por ejemplo, por 30 fonemas que equivalga a 7 palabras, entonces la probabilidad de que los 30 fonemas en la transcripción obtenida de la frase completa sean correctos, es menor que la probabilidad de que las 7 palabras de la transcripción obtenida sean correctas.

4.2.2. Reconocimiento basado en modelos de tri-fonemas.

En estas pruebas de reconocimiento a nivel de palabras, se utilizaron los modelos de tri-fonemas con 39 Gaussianas que se obtuvieron en la fase de entrenamiento (véase la *Sección 3.2.4 y 3.2.8.2*. Para el desarrollo de dichas pruebas se hizo uso de:

- una gramática a nivel de palabras, en la que todas las palabras tienen la misma probabilidad de aparición tras una palabra dada, como se comentó en la *Sección 3.2.7.1*,
- y de un diccionario que contiene el conjunto de palabras contenidas en la base de datos y sus correspondientes transcripciones a nivel de tri-fonemas. Dicho diccionario fue creado a partir del diccionario fonético de palabras (*Figura 30*) mediante las herramientas proporcionadas por HTK.

```
about       [about]       ah+b ah-b+aw b-aw+t aw-t sp
above       [above]       ah+b ah-b+ah b-ah+v ah-v sp
absorbed    [absorbed]    ah+b ah-b+z b-z+ao z-ao+r ao-r+b r-b+d b-d sp
account     [account]     ah+k ah-k+aw k-aw+n aw-n+t n-t sp
activate    [activate]    ae+k ae-k+t k-t+ah t-ah+v ah-v+ey v-ey+t ey-t sp
activity    [activity]    ae+k ae-k+t k-t+ih t-ih+v ih-v+ah v-ah+t ah-t+iy t-iy sp
actual      [actual]      ae+k ae-k+ch k-ch+ah ch-ah+w ah-w+ah w-ah+l ah-l sp
      .
      .
      .
```

Figura 32- Algunas entradas del diccionario a nivel de tri-fonético.

Los resultados obtenidos en dicha prueba (*Figura 33)*, muestran un *46.12 %* de palabras correctas y una precisión de *34.53 %*.

```
====================== HTK Results Analysis =======================
  Date: Mon Aug 18 10:34:45 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/triphones/words_test.mlf
  Rec : /media/datos/BD/LM/tri_recognition/210__60/recout_20_100.mlf
---------------------- Overall Results ---------------------------
SENT: %Correct=20.53 [H=1149, S=4448, N=5597]
WORD: %Corr=46.12, Acc=34.53 [H=12289, D=2857, S=11499, I=3089, N=26645]
====================================================================
```

Figura 33- Resultados obtenidos en reconocimiento a nivel de palabras con los modelos basados en tri-fonemas de 39 Gaussianas obtenidos.

Nótese que en esta prueba el porcentaje de palabras correctas, *46.12%*, también es menor que el de fonemas correctos de la prueba de reconocimiento fonético (*véase la Figura 29*) y que la precisión también es menor. La justificación es la misma que en el reconocimiento a nivel de palabras basado en modelos fonéticos de la *Sección 4.2.1*. No obstante, estos resultados son mejores que los obtenidos en dicho reconocimiento a nivel de palabras. Ello es debido a que en este reconocimiento se hace uso de los modelos de tri-fonemas, que fueron creados teniendo en cuenta las características de los parámetros según el contexto acústico, dotando al reconocedor de palabras de mayor precisión y robustez.

Nótese también, el porcentaje de frases completas correctas ha aumentado de un *16.60%* (en el reconocimiento de palabras basado en modelos fonéticos) a un *20.53%*. Ello es debido a la mejora del porcentaje de palabras y precisión con el uso de modelos basados en tri-fonemas.

4.2.3. Reconocimiento con N-best.

Para último se han desarrollado pruebas de reconocimiento de palabras teniendo en cuenta N-best hipótesis, con el fin de terminar de evaluar los modelos obtenidos y obtener los lattices correspondientes para los archivos de evaluación. Estos *lattices* serán utilizados en otros proyectos para diversas aplicaciones (*Word*-S*potting*, reconocimiento de idioma, etc), siguiendo esta nueva línea de investigación.

Los *lattices* obtenidos en el desarrollo de estas pruebas mediante las herramientas de HTK están en el formato de HTK llamado SLF (S*tandar Lattice Format*). Dicho formato es compatible con la mayoría de herramientas de tratamiento de *lattices*, y en concreto con el *lattice-tool*, utilizado en el grupo de investigación de la ATVS.

Estas pruebas se han llevado a cabo con los modelos de 39 Gaussianas, tanto con los modelos fonéticos, como con los modelos de tri-fonemas. Los resultados obtenidos son los siguientes:

- Para los modelos fonéticos:
  - Teniendo en cuenta las 100 mejores hipótesis (100-best).

```
====================== HTK Results Analysis =======================
  Date: Mon Jul 14 00:37:00 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/Ngramas/words_test.mlf
  Rec : /media/datos/BD/LM/ph_recognition/201__45/recout_20_100.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=33.73 [H=1888, S=3709, N=5597]
WORD: %Corr=60.40, Acc=49.58 [H=16094, D=2534, S=8017, I=2883, N=26645]
====================================================================
```

Figura 34- Resultados obtenidos en reconocimiento 100-best a nivel de palabras con los modelos basados en fonemas de 39 Gaussianas.

  - Teniendo en cuenta las 75 mejores hipótesis (75-best).

```
====================== HTK Results Analysis =======================
  Date: Thu Jul 17 21:38:25 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/Ngramas/words_test.mlf
  Rec : /media/datos/BD/LM/ph_recognition/201__45/recout_20_75.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=33.14 [H=1855, S=3742, N=5597]
WORD: %Corr=59.63, Acc=48.72 [H=15889, D=2581, S=8175, I=2908, N=26645]
====================================================================
```

Figura 35- Resultados obtenidos en reconocimiento 75-best a nivel de palabras con los modelos basados en fonemas de 39 Gaussianas.

- Teniendo en cuenta las 50 mejores hipótesis (50-best).

```
====================== HTK Results Analysis =======================
  Date: Sat Aug  2 21:02:47 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/Ngramas/words_test.mlf
  Rec : /media/datos/BD/LM/ph_recognition/201__45/recout_20_50.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=32.12 [H=1798, S=3799, N=5597]
WORD: %Corr=58.25, Acc=47.18 [H=15522, D=2634, S=8489, I=2951, N=26645]
====================================================================
```

Figura 36- Resultados obtenidos en reconocimiento 50-best a nivel de palabras con los modelos basados en fonemas de 39 Gaussianas.

- Teniendo en cuenta las 25 mejores hipótesis (25-best).

```
====================== HTK Results Analysis =======================
  Date: Wed Aug  6 18:01:42 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/Ngramas/words_test.mlf
  Rec : /media/datos/BD/LM/ph_recognition/201__45/recout_20_25.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=30.25 [H=1693, S=3904, N=5597]
WORD: %Corr=55.76, Acc=44.32 [H=14856, D=2733, S=9056, I=3047, N=26645]
====================================================================
```

Figura 37- Resultados obtenidos en reconocimiento 25-best a nivel de palabras con los modelos basados en fonemas de 39 Gaussianas.

- Para los modelos de tri-fonemas:
  - Teniendo en cuenta las 100 mejores hipótesis (100-best).

```
====================== HTK Results Analysis =======================
  Date: Mon Aug 18 10:34:30 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/triphones/words_test.mlf
  Rec : /media/datos/BD/LM/tri_recognition/210__60/recout_20_100.mlf
---------------------- Overall Results ----------------------
SENT: %Correct=41.83 [H=2341, S=3256, N=5597]
WORD: %Corr=67.73, Acc=57.93 [H=18047, D=2181, S=6417, I=2611, N=26645]
====================================================================
```

Figura 38- Resultados obtenidos en reconocimiento 100-best a nivel de palabras con los modelos basados en tri-fonemas de 39 Gaussianas.

- Teniendo en cuenta las 75 mejores hipótesis (75-best).

```
====================== HTK Results Analysis =======================
  Date: Mon Aug 18 10:34:58 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/triphones/words_test.mlf
  Rec : /media/datos/BD/LM/tri_recognition/210__60/recout_20_75.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=40.93 [H=2291, S=3306, N=5597]
WORD: %Corr=66.96, Acc=57.08 [H=17842, D=2230, S=6573, I=2634, N=26645]
===================================================================
```

Figura 39- Resultados obtenidos en reconocimiento 75-best a nivel de palabras con los modelos basados en tri-fonemas de 39 Gaussianas.

- Teniendo en cuenta las 50 mejores hipótesis (50-best).

```
====================== HTK Results Analysis =======================
  Date: Mon Aug 18 10:35:18 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/triphones/words_test.mlf
  Rec : /media/datos/BD/LM/tri_recognition/210__60/recout_20_50.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=39.61 [H=2217, S=3380, N=5597]
WORD: %Corr=65.80, Acc=55.77 [H=17532, D=2296, S=6817, I=2671, N=26645]
===================================================================
```

Figura 40- Resultados obtenidos en reconocimiento 50-best a nivel de palabras con los modelos basados en tri-fonemas de 39 Gaussianas.

- Teniendo en cuenta las 25 mejores hipótesis (25-best).

```
====================== HTK Results Analysis =======================
  Date: Mon Aug 18 10:35:31 2008
  Ref : /home/voz/sergio/Bases_datos/SPEECHDAT/triphones/words_test.mlf
  Rec : /media/datos/BD/LM/tri_recognition/210__60/recout_20_25.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=37.09 [H=2076, S=3521, N=5597]
WORD: %Corr=63.45, Acc=53.16 [H=16906, D=2438, S=7301, I=2741, N=26645]
===================================================================
```

Figura 41- Resultados obtenidos en reconocimiento 25-best a nivel de palabras con los modelos basados en tri-fonemas de 39 Gaussianas.

Nótese, que tanto el porcentaje de palabras correctas y la precisión, como el porcentaje de frases completas correctas es mayor en cualquiera de los reconocimientos *n*-best que hemos realizado que en los resultados mostrados en las *Secciones 4.2.1 y 4.2.2*. Esto significa que la correcta transcripción de algunos archivos de evaluación, para los que se ha obtenido una transcripción errónea, se encuentra entre las *n* hipótesis más probables. Obviamente al aumentar el número de las hipótesis más probables a tener en cuenta, *n*, los resultados obtenidos mejoran hasta llegar a un límite. Observe como del reconocimiento 75-best al 100-best los resultados aumentan menos de un *1%* mientras que del reconocimiento 25-best al 50-best, los resultados mejoran entre en un *2-3%.*

En la siguiente tabla se muestra una comparativa de los resultados obtenidos en las pruebas de reconocimiento de palabras n-best, en la que se puede observar la notable mejora de los resultados.

		%Corr	%Acc	%Corr - frases
Reconocimiento de palabras - modelos de fonemas	1-best	38.36	25.18	16.60
	25-best	55.76	44.32	30.25
	50-best	58.25	47.18	32.12
	75-best	59.63	48.72	33.14
	100-best	60.40	49.58	33.73
Reconocimiento de palabras - modelos de tri-fonemas	1-best	46.12	34.53	20.53
	25-best	63.45	53.16	37.09
	50-best	65.80	55.77	39.61
	75-best	66.96	57.08	40.93
	100-best	67.73	57.93	41,83

Tabla 10- Resultados obtenidos en reconocimiento de palabras 1, 25 50, 75, 100-best.

A la vista de la comparativa de resultados, mostrada en la *Tabla 10*, nótese la gran mejora de los resultados que supone pasar del reconocimiento con 1-best reconocimiento 25-best. Para el reconocimiento basado de modelos fonéticos se han obtenido las siguientes mejoras:

- Una mejora de un 17.40% en el porcentaje de palabras correctas.
- Un 19.14% de mejora en la precisión del sistema.
- Un 13.45% en el porcentaje de frases completas correctas.

Y para el reconocimiento basado en modelos dependientes del contexto, las mejoras obtenidas son:

- Un 17.33% en el porcentaje de palabras correctas.
- Un 18.63% de mejora en la precisión del sistema.
- Un 16.62% en el porcentaje de frases completas correctas.

Nótese también, que para el resto de los aumentos del número de hipótesis a tener en cuenta (50, 75 y 100-best), las mejoras obtenidas no son tan significativas y disminuyen según se aumenta el número de hipótesis, lo cual supone un gran aumento del tiempo de computación. Por lo tanto, tomaremos los resultados proporcionados por el reconocimiento con 25-best como los resultados óptimos de este experimento.

4.3. Otros experimentos.

Aparte de los experimentos anteriores, también se intentó el desarrollo de otros, que son descritos a continuación:

- Pruebas de reconocimiento con tri-fonemas, con el fin de observar los resultados y mejoras obtenidas respecto al reconocimiento fonético. Para dicho reconocimiento se requiere una gramática a nivel de tri-fonemas. Pero en la creación de dicha gramática mediante HTK se tuvieron problemas de memoria dinámica, debido a la gran cuantía de fonemas (incluidos los tri-fonemas y bi-fonemas) a incluir en la gramática, 62413.

- Pruebas de reconocimiento haciendo uso de una gramática basada en un modelo de lenguaje, con el fin de obtener mejoras en el reconocimiento de palabras. Para ello, se creó un modelo lenguaje basado en N-gramas (con *N = 1, 2, 3*) mediante las herramientas proporcionadas por HTK. Dicho modelo se entrenó con "*English Gigaword, 2ª edición*". Dicha base de datos fue creada por el *Linguistic Data Consortium (LDC)*, recopilando textos de noticias en inglés durante varios años.

  Finalmente, el modelo de lenguaje obtenido basado en tri-gramas (*N=3*) no se pudo aplicar al reconocimiento, ya que HTK no soporta dicha operación. Y aunque, sí que soporta el reconocimiento usando gramáticas basadas en modelos de lenguaje de bi-gramas (*N=2*), tampoco se consiguió dicho reconocimiento, por problemas con la propia herramienta que no se pudieron solventar.

  Una opción, hubiese sido utilizar *Sphinx*, tanto para el entrenamiento del modelo de lenguaje, como para el reconocimiento y obtención de los resultados. Pero para ello, habría que realizar todo el proceso de entrenamiento de los modelos acústicos con *Sphinx*, o bien, cambiar las transcripciones de los archivos de evaluación y los modelos acústicos entrenados al formato utilizado por *Sphinx*. Esta última opción, que parecería la más apropiada en un caso como el nuestro, en el que ya hemos realizado el entrenamiento de los modelos acústicos con HTK, no es nada trivial, ya que la transformación del formato de HTK al de *Sphinx* es complicada, especialmente en el caso de tri-gramas.

5. Conclusiones y trabajo futuro

5.1. Conclusiones

Durante el desarrollo del proyecto y a la vista de los resultados obtenidos, extraemos las siguientes conclusiones:

- El aumento del número de Gaussianas para el modelado de los estados de los distintos modelos (de fonemas o tri-fonemas) mejora los resultados en términos, tanto de porcentaje de fonemas/palabras y frases correctas, como de precisión del sistema. De esta manera se obtiene, por tanto, un sistema de mayor exactitud y mayor tasa de aciertos sin necesidad de inserciones espurias.

- Se ha conseguido obtener unos modelos fonéticos que proporcionan unos buenos resultados en el reconocimiento fonético en términos de porcentaje de fonemas correctos y de precisión del sistema, aunque con un bajo porcentaje de frases completas correctas.

- El desarrollo de reconocimiento a nivel de palabras mejora el porcentaje de frases completas correctas, debido a la utilización de una gramática sin consideraciones léxicas-semánticas de la lengua. Por tanto, dichos resultados se podrían mejorar con la construcción y utilización de una gramática que modelase mejor el idioma.

- Mediante la creación de modelos basados en tri-fonemas y la atadura de diversos estados se aprovechan las características de los parámetros para un fonema según su contexto acústico. De esta manera, como hemos podido observar en los experimentos realizados, se obtienen mejores resultados tanto en términos de porcentaje de palabras y frases correctas, como en términos de precisión del sistema.

- Como se ha analizado en la *Sección 4.2.3.*, teniendo en cuenta sólo las 25 hipótesis más probables en el reconocimiento de palabras obtiene un incremento en torno al 17-20%, tanto en el porcentaje de palabras y frases correctas, como en la precisión del sistema. Por lo que, sería interesante aprovechar el estudio de dichas hipótesis para re-estimar los modelos e intentar conseguir dicha mejora.

5.2. Trabajo futuro

Dados los experimentos realizados y las conclusiones extraídas sería interesante:

- Construir una gramática basada en tri-fonemas y realizar un reconocimiento con tri-fonemas con los datos de evaluación de la base de datos, con el fin de observar los resultados y las mejoras obtenidas respecto al reconocimiento fonético.

- Construir una gramática a nivel de palabras basada en un modelo de idioma, teniendo en cuenta los aspectos léxicos-semánticos de la lengua, y mejorar, así, los resultados obtenidos en el reconocimiento a nivel de palabras.

- Utilizar los *lattices* que se obtienen en el reconocimiento fonético basado en N-best, para aplicaciones de *Word-Spotting*. Un reconocedor fonético genera siempre las secuencias de fonemas reconocidas, pero en el proceso de búsqueda evalúa otras muchas opciones que pueden ser representadas en *lattices*, como ya se ha comentado. Entonces, la técnica a utilizar sería hacer un reconocimiento fonético para obtener un *lattice* con las secuencias de fonemas más probables y buscar las palabras clave en el *lattice*. De esta manera, una vez realizada la indexación, se puede buscar cualquier palabra, ya que se buscan secuencias de fonemas. Con ello, se consigue que la aplicación no esté limitada a un vocabulario finito y, además, las búsquedas en el *lattice* son bastantes rápidas.

Referencias

L. R. Bahl, F. Jelinek. *Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition*. *IEEE Trans. Informat. Theory*, vol. IT-21, pp. 404-411, 1975.

L. R. Bahl, F. Jelinek, R. L. Mercer. *A maximum likelihood approach to continuous speech recognition*. *IEEE Trans. Patern Anal. Machine Intell.*, vol. PAMI-5, pp.179-190, 1983.

J. K. Baker. *The dragon system–An Overview*. *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-23, no. 1, pp.24-29, Feb. 1975.

R.Bakis. *Continuous speech word recognition cia centi-second acoustic states*. *In Proc. ASA Meeting* (Washington, DC), Abr. 1976.

L. E. Baum. *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*. Inequalities, vol. 3, pp. 1-8, 1972

L.E. Baum, J. A. Egon. *An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology*. *Bull. Amer. Meteorol. Soc*., vol.73, pp. 360-363, 1997.

L. E. Baum, T. Petrie. *Statistical inference for probabilistic functions of finite state Markov chains*. *Ann. Math. Stat.*, vol.37, pp. 1554-1563, 1996.

L. E. Baum, T. Petrie, G. Soules, N. Weiss. *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*. *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164-171, 1970.

L.E. Baum, G. R. Sell. *Growth functions for transformations on manifolds*. *Pac. J. Math.*, vol. 27, no, 2, pp. 211-227, 1968.

J. Campbell. *Speaker Recognition: A Tutorial. Proc. of the IEEE*, 1997, **85**(9), pp. 1437-1462.

F.Casacubierta, E. Vidal. *Reconocimiento automático del habla.* Marcombo, 1987.

F.Casacubierta, E. Vidal. *Reconocimiento automático del habla: Metodologías y arquitecturas*. En Inteligencia artificial: Conceptos, métodos y aplicaiones, Marcombo, 1987.

The CMU Pronouncing Dictionary, disponible en http://www.speech.cs.cmu.edu/cgi-bin/cmudict

S. Davis, P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. *IEEE Trans*., ASSP 28 , pp. 357-366, 1980.

A. P. Dempster, N. M. Laird, D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. *J. Roy Stat. Soc.*, vol. 39, no.1, pp. 1-38, 1977.

R. O. Duda, P. E. Hart, D. G. Store. *Patter Classification*. Wiley, 2001.

Gil, Juana. *Los sonidos del lenguaje*. Madrid, Síntesis, 1989.

M. Helander. *Handbook of Human-Computer Interaction*. Amsterdam,North-Holland, 1997.

F. Jelinek. *A fast sequential decoding algorithm using a stack*. *IBM J. Res. Develop.*, vol.13, pp.675-685, 1969.

F. Jelinek. *Continuous speech recognition by statistical methods*. *Proc. IEEE*, vol. 64, pp. 532-536, Apl. 1976.

F. Jelinek, L. R. Bahl, R. L. Mercer. *Design of a linguistic statistical decoder for the recognition of continuous speech*. *IEEE Trans. Informat. Theory*, vol. IT 21, pp.250-256, 1975.

F. Jelinek, L. R. Bahl, R. L. Mercer. *Continuous speech recognition: Statistical methods*. *In Handbook of Statistics, II*, P. R. Krishnaiad, Ed. Amsterdam, The Netherlands: Nort-Holland, 1982.

J. C. Junqua and J.P. Haton. *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers. 1996.

Llisterri, Joaquim. Introducción a la fonética: el método experimental. Barcelona, Anthropos, 1991.

B. C. J. Moore. *An introduction to the psychology of hearing*. *Fifth edition*., Academic Press, San Diego. Enero, 2003.

N. Morales Mombiela. *Robust Speech Recognition Under Band-Limited Channels And Other Channel Distortions*. PhD Thesis, pp 5-7. Junio 2007.

L. Nguyen, R. Schwartz, Y. Zhao and G. Zavaliagkos (1994). *Is N-best dead? Proceedings workshop on Human Language Technology*. pp. 411-414. Marzo, 1994.

H. Nyquist (1928). *Certain topics in telegraph transmission theory*. *Reprinted in Proceedings of the IEEE*. vol. 90, pp. 280-305. Febrero, 2002.

D. O'Shaughnessy. *Invited Paper: Automatic speech recognition: History, methods and challenges*. PR(41), No.10, pp.2965-2979. Octubre 2008.

L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. *In Proceedings of the IEEE*, vol. 77, nº 2, pp. 257-286. Febrero 1989.

F. Richardson, M. Ostendorf and J.R. Rohlicek. *Lattice-based search strategies for large vocabulary speech recognition*. *Proceedings ICASSP'95*, vol. 1, pp. 576-579. Mayo, 1995.

C. Shanon. *Communication in the presence of noise*. *Proceedings of the IRE*, vol. 37, pp. 10-21. *Reprinted in Proceedings of IEEE*, vol. 86, pp. 447-457. Enero, 1949.

N. S. Trubetzkoy. Principios de fonología. Madrid, Cincel. Varias ediciones, 1939

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland. *The HTK Book.* Versión 3.4, 2006. Disponible en http://htk.eng.cam.ac.uk/.

Glosario

DCT (*Discrete Cosine Transform*)
Transformada de coseno discreta.

DTW (*Dynamic Time Warping*)
Alineamiento Temporal Dinámico.

English Gigaword
Base de datos de textos de noticias creada por *Lingistic Data Consortium*.

Escala de frecuencia Bark
Clase de escala de banda crítica creada con objeto de aproximarse a la sensibilidad del oído humano.

Escala de frecuencia Mel
Clase de escala de banda crítica creada con objeto de aproximarse a la sensibilidad del oído humano.

FFT (*Fast Forier Transform*)
Transformada rápida de Fourier.

HMM (*Hidden Markov Model*)
Modelo Oculto de Markov.

HTK (*Hidden Markov Model Toolkit*)
Conjunto de herramientas para la creación y tratamiento de HMMs, diseñado por la universidad de Cambridge.

IPA (*International Phonemes Association*)
Representación estandarizada de los sonidos del habla.

Lattice
Representación compacta de una serie de hipotéticas posibles transcripciones para un archivo de audio concreto. Es muy utilizado en sistemas de reconocimiento de habla.

LM (*Language model*)
Modelo de lenguaje.

LDC (*Lingistic Data Consortium*)
Consorcio abierto de universidades, compañias y laboratorios de investigación guernamentales, que colecta y distribuye bases de datos de texto y habla, léxicos y otras fuentes para la investigación.

MFCCs (*Mel-Frequency Cepstral Coefficients*)
Coeficientes cepstrales en escala de frecuencias Mel.

MLF (*Master Label Format*)
Formato de HTK para archivos que contienen transcripciones.

NIST (*National Institute of Standars and Technology*)
Organismo federal, no regulador, perteneciente a la Cámara de Comercio de los Estados Unidos que desarrolla y promueve medidas, estándares y tecnología para aumentar la productividad, facilitar el comercio y mejorar la calidad de vida.

RAH
Reconocimiento Automática de Habla.

Reconocimiento del habla
Tecnología que permite que una máquina reconozca las palabras pronunciadas. El reconocimiento del habla no es una tecnología biométrica.

Reestimación Baum-Welch
Método de entrenamiento de un Modelo Oculto de Markov.

SLF (*Estándar Lattice Format*)
Formato de HTK para archivos que contienen *lattices.*

Sphinx
Conjunto de herramientas para el tratamiento de voz diseñado por la universidad de Carnegie-Mellon

SPL (*Sound Pressure Level*)
Medida del nivel de presión de un sonido. Se mide en dB.

VQ (*Vector Quantization*)
Cuantificación Vectorial

Word- Spotting es una nueva línea estratégica de I+D, que consiste en determinar si una palabra clave aparece en una grabación y en que instante de tiempo.

PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (*Software* incluido)..... 2.000 €
- Alquiler de impresora láser durante 11 meses ..... 100 €
- Material de oficina ..... 200 €
- Total de ejecución material ..... 2.300 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 368 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 138 €

4) Honorarios Proyecto

- 11 meses a 475 € / mes ..... 5225 €

5) Material fungible

- Gastos de impresión..... 120 €
- Encuadernación..... 10 €

6) Subtotal del presupuesto

- Subtotal Presupuesto..... 7655 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto ..... 1224.8 €

8) Total presupuesto

- Total Presupuesto..... 8879,8 €

Madrid, Septiembre de 2008

El Ingeniero Jefe de Proyecto

Fdo.: Sergio Lucas Bermejo
Ingeniero de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un SISTEMA DE TRANSCRIPCION AUTOMATICA DE VOZ ESPONTANEA PARA RECUPERACIÓN DE INFORMACION. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es

obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.