TESIS PUCP


PONTIFICIA **UNIVERSIDAD CATÓLICA** DEL PERÚ

Esta obra ha sido publicada bajo la licencia Creative Commons
Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5 Perú.
Para ver una copia de dicha licencia, visite
http://creativecommons.org/licenses/by-nc-sa/2.5/pe/


PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

TESIS PUCP

# PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## FACULTAD DE CIENCIAS E INGENIERÍA

# MODULO DE RECONOCIMIENTO DE VOZ A TEXTO INDEPENDIENTE DE LOCUTOR PARA SISTEMAS DE DIALOGO

Tesis para optar el Título de **Ingeniero Electrónico**, que presenta el bachiller:

**Ursula Del Milagro García García**

**ASESOR:** **Ing. Donato Andrés Flores Espinoza**

Lima, septiembre del 2009

SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## RESUMEN

En la actualidad, gracias al avance de la tecnología y del desarrollo matemático, los sistemas que interactúan con el usuario utilizando el habla son una realidad en varias partes del mundo, principalmente dando información sobre diversos dominios, ya sean viajes en avión y en tren, búsquedas medicas, etc. Sin embargo, construir uno implica una investigación multidisciplinaria, donde se mezclan la lingüística, el procesamiento digital de señales y la inteligencia artificial.

La primera etapa de este sistema es la etapa del reconocimiento de habla, la cual recoge la información dada por el usuario, la cual va a ser interpretada. Por ello, su diseño debe ser cuidadoso, procurando equilibrar la rapidez, la robustez frente al ruido y la precisión, para lo cual son necesarias muchas pruebas e implementación de algoritmos que mejoren estos tres aspectos.

Este trabajo presenta el primer bosquejo de la etapa de reconocimiento de habla, y se enmarca en el proyecto DAI “Sistema de Dialogo Hablado Aplicado a una Recepcionista Telefónica Automática” del año 2004, y busca desarrollar un modulo basado en unidades inferiores a la palabra que sea independiente de locutor; es decir, que pueda ser usado por la mayor cantidad de personas posible, con un porcentaje de errores muy parecido. Se presenta el desarrollo en software utilizando lenguaje C en entorno Windows, con un conjunto de herramientas llamado HTK, desarrollado por la Universidad de Cambridge, Inglaterra.

SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

¡Ja! Ya me estoy imaginando un libro del Mundodisco escrito con software de reconocimiento de voz. ¡Sobre todo en esta casa infestada de gatos! ´¡Se escribe Ankh-Morpork, maquina tonta! ¡Bajate de la mesa!´

Terry Pratchett

SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

A mi familia (Juan, Delia y Jessica) por el cariño recibido en todos estos años.

A mis amigos de la universidad y de la arqueria, por todo lo compartido.

Al GPDSI, por las largas horas frente a la computadora en el laboratorio.

Al Ing. Andres Flores, por el apoyo brindado.

SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## INDICE



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## INDICE DE FIGURAS



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## INDICE DE TABLAS



SOME RIGHTS RESERVED

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

# 1 INTRODUCCIÓN



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Uno de los anhelos del ser humano han sido las maquinas inteligentes, con las que pueda interactuar como con otro ser humano, utilizando la forma mas natural y antigua de comunicación : el lenguaje. El avance de la tecnología ha permitido avanzar a grandes pasos hacia esta meta; actualmente, existen robots que son controlados por voz, ya sea asistentes o juguetes, y sistemas que dialogan de manera mas o menos natural con el usuario.

Una parte importante de estos sistemas es la etapa de reconocimiento de habla. Esta etapa modela el sistema auditivo humano y parte del procesamiento cerebral del lenguaje, sin llegar a la comprensión de lo que dice el hablante. La precisión en esta etapa es importante, puesto que su salida es la entrada a la etapa de comprensión, donde se interpreta lo que el hablante quiere decir, y se genera la respuesta adecuada. Por esta razón, generalmente se privilegia la precisión sobre la rapidez y la robustez frente al ruido, aunque en los últimos años han surgido nuevas técnicas matemáticas que dan un rendimiento bastante apreciable en estas dos ultimas características.

Hay distintos métodos para diseñar esta etapa, los cuales se pueden clasificar en dos grupos : métodos paramétricos y métodos estocásticos. Estos últimos, basados en probabilidades, son los que se usan con mayor frecuencia, y han dado buenos resultados. Dentro de este grupo destacan los modelos ocultos de Markov (HMM – Hidden Markov Models), los cuales se han venido utilizando desde hace varios años en este tipo de aplicación. En los últimos años, con el desarrollo de las redes neuronales, la tendencia es de mezclar los HMM con MLP (MultiLayer Perceptron), lo cual ha aumentado el rendimiento y la flexibilidad de los reconocedores.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

En este trabajo, se busca realizar el primer bosquejo de una etapa de reconocimiento de habla orientada a un sistema de dialogo basado en unidades inferiores a la palabra, por lo cual debe funcionar para varias personas, sin importar el sexo y la edad. Esto es relativo, ya que siempre funcionara mejor con las personas que presenten características similares a aquellas que prestaron su voz para el entrenamiento. También debe ser lo mas preciso posible, para lo cual se identificaran los factores mas influyentes con vista a mejorarlos.

La organización de este trabajo es la siguiente : en el capitulo 2, presentamos la descripción del problema, revisando la situación actual de los sistemas de dialogo y de reconocimiento de voz como parte de ellos; en el capitulo 3 se introducen las técnicas matemáticas y su implementación en lenguaje C, por medio del conjunto de herramientas HTK; en el capitulo 4 se presentan los resultados de las pruebas a las que fue sometido el sistema; en el capitulo 5 se enumeran las observaciones y conclusiones que se extraen de dichas pruebas, y en el capitulo 6 se presentan las recomendaciones con vista a mejorar este primer bosquejo.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

# 2 ANÁLISIS DEL PROBLEMA



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

En este primer capítulo, se analiza el contexto en que está ubicado el presente trabajo, de los sistemas de diálogo y su empleo en el mundo actual. Luego se hará un repaso al reconocimiento automático de voz, sus características y los problemas a los que se enfrenta, para terminar presentando las proyecciones que se desean cumplir con este trabajo.

## *2.1 Sistemas de diálogo*

Con el desarrollo de la tecnología, el concepto de un sistema que interaccione con el usuario utilizando el habla se hace cada vez más factible. Es por eso que el desarrollo e implementación de sistemas de diálogo hablado ha cobrado auge en el mundo, dentro de dominios específicos.

Un sistema de diálogo hablado es un sistema de interacción hombre-máquina, que se vale del habla, como forma más natural de comunicación, para requerir y transmitir información a un usuario, dentro de un determinado contexto, y por lo general orientado a tareas de atención al cliente, control de dispositivos y acceso a la información.

En la actualidad, hay varios sistemas de diálogo en uso alrededor del mundo, los cuales emplean incluso más de un idioma, tal como se muestra en la tabla 2.1.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

| NOMBRE | PAÍS / IDIOMA | OBJETIVO |
|---|---|---|
| PEGASUS | Estados Unidos / Inglés | Proveer información sobre vuelos a través de una línea telefónica. Puede responder preguntas sobre horarios de partida y llegada o realizar un plan de vuelo para el día en que el usuario lo requiera. |
| HOMEY | Inglaterra / Inglés | Interface hablada para sistemas de información clínica. |
| LODESTAR | China / Mandarín | Brindar información turística y planea itinerarios de acuerdo a los requerimientos del usuario. |
| OVIS | Holanda / Holandés | Brindar información sobre el transporte público. |
| AthosMail | Europa / Multilingue (Finlandés, inglés, sueco) | Permitir al usuario acceder a su buzón de correo utilizando el teléfono. |
| CLARISSA | Estados Unidos / Inglés | Permitir el control de una nave espacial por medio del diálogo. |
| PENATES | Estados Unidos / Inglés | Brindar información sobre aproximadamente 1000 restaurantes en más de 100 ciudades, principalmente Boston, Cambridge, Brookline y Somerville. |

**Tabla 2-1 Algunos sistemas de diálogo hablado en el mundo [11-16]**

A los sistemas mencionados se agrega el que esta siendo desarrollado actualmente por la Pontificia Universidad Católica del Perú, el cual se orienta a dar información sobre cines mediante el uso de una línea telefónica, del cual es parte este trabajo.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

### 2.1.1 Partes de un sistema de diálogo

Un sistema de diálogo está compuesto por 4 módulos :

- Módulo de Reconocimiento de Habla Continua (Continuous Speech Recognition Module - CSR), encargado de reconocer las elocuciones del usuario, procesarlas y dar como resultado una cadena de texto que contenga dicha entrada hablada. Suele utilizar patrones estadísticos (HMM), modelos de lenguaje (n-grams) y un lexicon para el reconocimiento.
- Módulo de Procesamiento de Lenguaje Natural (Natural Language Processing Module - NLP), toma la cadena de texto que le entrega el CRS y decide qué es lo que el usuario ha querido decir, buscando secuencias de concepto en dicha cadena.
- Módulo de Manejo de Diálogo (Dialog Management – DM), genera la respuesta del sistema frente a lo expresado por el usuario, ya sea para pedir más información, haciendo una pregunta de clarificación, o dando la información requerida. Esta respuesta es otra cadena de texto.
- Módulo de Texto-a-Habla (Text-To-Speech - TTS), convierte la respuesta generada por el DM en sonido, por medio de técnicas de síntesis de voz.

La relación entre estos módulos se puede ver en la figura 2.1.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

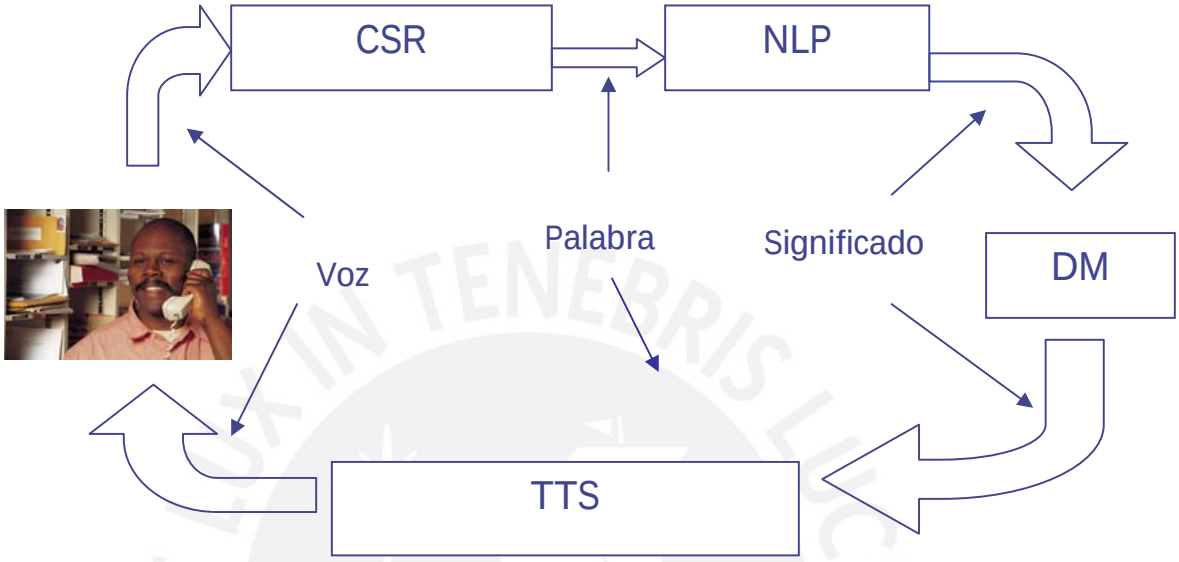



**Figura 2-1 Partes de un sistema de diálogo**

En este trabajo se presenta el inicio del desarrollo del primer módulo: Reconocimiento de Habla Continua, por lo que profundizaremos en él, presentando los requerimientos deseados y los problemas a enfrentar.

## *2.2 Módulo de Reconocimiento de Habla Continua (CSR)*

Como se dijo anteriormente, el objetivo de este módulo es captar adecuadamente el habla emitida por el usuario y traducirla a una cadena de texto, la cual será interpretada por el módulo siguiente. Antes de describir las técnicas matemáticas que hacen posible esto, introduciremos brevemente algunos conceptos sobre los sistemas de reconocimiento automático de habla (Automatic Speech Recognition – ASR).



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

### 2.2.1 Reconocimiento automático de habla

El reconocimiento automático de habla es la habilidad de las máquinas para “entender” lo que, a lo menos en teoría, cualquier persona dice en un lenguaje determinado. Para esto, emplea modelos de los sonidos propios de un idioma, además de un conjunto de reglas gramaticales y variedades de pronunciación para decodificar el habla; es decir, realiza una comparación de patrones (modelos estocásticos) para obtener un rango de probabilidades, en el cual se basa para escoger el modelo con la probabilidad más alta.

Estos patrones necesitan ser entrenados concienzudamente para que el sistema funcione correctamente; es por ello que los sistemas que emplean reconocimiento de habla por lo general necesitan abundante material de entrenamiento y sus algoritmos pueden tardar tiempo en ejecutarse.

#### 2.2.1.1 Características

Los sistemas de reconocimiento de habla se clasifican en varios tipos, de acuerdo a como satisfagan las siguientes condiciones :

- Número de personas para las que el sistema debe funcionar

  Si se desea que el sistema funcione para una única persona, o *locutor*, se dice que es *dependiente de locutor*, puesto que sólo se le entrena con esa persona. Si al contrario, se desea que el sistema funcione para la mayor cantidad posible de usuarios, será *independiente de locutor*, y deberá ser entrenado con una gran cantidad de locutores;



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

es decir, que el corpus o conjunto de voces utilizado deberá ser lo más grande y variado posible.

Sin embargo, una desventaja de este tipo de sistemas es que su mejor funcionamiento se dará cuando lo usen las personas con cuyas voces se ha entrenado, o cuando lo usen personas con características similares a estas (edad, lugar de procedencia, estrato social, etc).

- Tamaño del vocabulario que manejan

Un sistema de reconocimiento puede ser de *vocabulario pequeño*, si no maneja más de 100 palabras; de *vocabulario mediano*, si maneja de 100 a 999 palabras, y de *gran vocabulario*, si sobrepasa las 1000 palabras.

El tamaño del vocabulario determina la elección de las unidades lingüísticas mínimas que usará el sistema. Si el vocabulario es pequeño, se usa como unidad lingüística la palabra; si el vocabulario es mediano, suele utilizarse también la sílaba o los fonemas, y si el vocabulario es muy extenso, se prefieren usar fonemas o fonos, ya que su número es más reducido y fácil de manejar por el sistema.

- Velocidad de las emisiones de los usuarios

Si los usuarios del sistema hablan haciendo pausas marcadas entre palabras (típicamente de 200 milisegundos), el sistema es de *palabras aisladas*, mientras que si se espera que los usuarios hablen a una velocidad normal, el sistema se diseñará para que reconozca *habla continua*.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

En la actualidad , los sistemas de alta eficiencia son los siguientes :

- Sistema de Vocabulario Pequeño.
- Sistemas de Vocabulario Grande pero de Palabra Aislada.
- Sistemas de Habla Continua pero delimitados por un contexto específico.

### 2.2.1.2 Condiciones adversas al reconocimiento de habla [4]

Existen tres condiciones que deben enfrentarse al diseñar un sistema de reconocimiento de habla, las cuales, de no ser adecuadamente consideradas, merman considerablemente su desempeño en un ambiente real.

a) Ruido ambiental

El ruido siempre ha sido un problema presente en el Procesamiento Digital de Señales, ya que, al mezclarse con la señal a analizar, la altera, e incluso puede llegar a anularla. Es por ello que una de las primeras preocupaciones al diseñar un sistema de reconocimiento es tomar en cuenta los niveles de ruido ambiental.

Las fuentes de ruido son numerosas. Por ejemplo, en una oficina, dichas fuentes incluyen computadoras, impresoras, máquinas de escribir, sonidos telefónicos y la conversación de otras personas. Tal como podemos ver en la tabla 2.2, el nivel de presión de sonido (Sound Pressure Level – SPL) en una oficina personal es alrededor de los 45 a 50 dB; en una oficina donde hay secretarias, el SPL puede



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

ser hasta 15 – 20 dB más alto, y en la cabina de un avión, SPLs de alrededor de 90dB o mayores son normales. A este nivel, el habla es difícilmente entendible para un humano, y más aún para una máquina.

| Sonidos comunes | Nivel de ruido (dB) | Efecto |
|---|---|---|
| Plataforma de lanzamiento de cohetes (sin protección auditiva) | 180 | Pérdida irreversible de la audición |
| Sirena de ataque aéreo | 140 | Ruido doloroso |
| Trueno | 130 | |
| Despegue de un jet (200 pies) | 120 | Esfuerzo vocal máximo |
| Concierto de rock | 110 | Extremadamente ruidoso |
| Camiones de basura.<br>Fuegos artificiales. | 100 | Muy ruidoso. |
| Camiones pesados (50 pies).<br>Tráfico urbano. | 90 | Muy fastidioso.<br>Daño auditivo (8 horas) |
| Alarma del despertador (2 pies).<br>Secador de pelo. | 80 | Fastidioso. |
| Restaurante ruidoso.<br>Oficina de negocios. | 70 | Dificultad al usar el teléfono |
| Unidad de aire acondicionado. | 60 | Intrusivo |



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

| | | |
|---|---|---|
| Conversaciones. | | |
| Tráfico moderado | 50 | Silencioso. |
| Sala de estar.<br>Dormitorio.<br>Oficina tranquila | 40 | |
| Biblioteca.<br>Susurros suaves. | 30 | Muy silencioso. |
| Estudio de grabación | 20 | |
| | 10 | Casi inaudible |
| | 0 | Umbral de la audición |

**Tabla 2-2 Niveles de sonido y respuesta humana (Extraído del website de Noise Pollution Clearinghouse)**

b) Distorsión

Antes de que el habla sea grabada y procesada para ser reconocida, sufre una serie de distorsiones espectrales, que se deben a distintas causas, tales como :

- El cuarto donde se está registrando la señal, ya que su nivel de reverberación puede alterar el espectro de ésta.
- El micrófono, que introduce distorsión dependiendo de su tipo y de su posición frente al hablante. Cuando el micrófono utilizado durante la grabación de muestras para el entrenamiento es diferente a aquel utilizado durante las pruebas del sistema, la precisión del sistema puede disminuir, ya que la distorsión espectral se convierte en un problema bastante serio.
- La red telefónica, si es que el sistema se ha diseñado para utilizarse en ella, ya que su ancho de banda se ha limitado para facilitar la transmisión de las señales, a



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

aproximadamente 200 – 3200 Hz, lo cual produce a la señal de voz una atenuación espectral de aproximadamente 10 dB.

c) Efectos articulatorios

Existen muchos factores que afectan la manera en que habla una persona, tales como enfermedades, estado de ánimo, interlocutor, etc. Los cambios característicos en la articulación del habla debidos a la influencia ambiental se conocen como el efecto Lombard, y pueden ser muy acentuados. Por ejemplo, se ha reportado que cuando el ambiente tiene un nivel de ruido de alrededor de 90 dB, la primera formante de una vocal incrementa su potencia, mientras que la segunda formante decae, lo cual puede devenir en un cambio apreciable en el espectro de dicha vocal.

El problema con el efecto Lombard es que aún no se comprende del todo cómo cuantificarlo, debido a la gran y rápida variación espectral que presenta, puesto que es inherente al proceso de producción del habla e independiente del contexto.

## *2.3 Proyecciones del trabajo presentado*

Dado todo lo anterior, se concluye que el Módulo de Reconocimiento de Habla de un sistema de diálogo deberá tener las siguientes características :

- Independencia de locutor, ya que el sistema está orientado a ser utilizado por varias personas.
- Vocabulario mediano o grande, dependiendo de la aplicación a la que vaya orientado; si está destinado a controlar algo, como un robot [1], le bastará un vocabulario mediano, pero si está orientado a dar información al cliente,



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

probablemente necesitará un vocabulario más extenso para poder abarcar, en lo posible, las múltiples variaciones del discurso hablado.

- Habla continua, puesto que es la velocidad normal en que los seres humanos se comunican. No obstante, esto agrega complejidad al sistema, porque no existe una velocidad fija de habla.
- Robustez frente al ruido y a la distorsión, lo cual se logra con el uso de algoritmos tales como filtros adaptativos y arreglos especiales de transductores.

Como con este trabajo se pretende dar un punto de partida al desarrollo del módulo de reconocimiento, se han simplificado estas condiciones, hasta llegar a lo siguiente :

- La independencia de locutor será relativa, ya que sólo un grupo pequeño de usuarios podrá ser reconocido de manera satisfactoria por el sistema. Esto se debe a que se utilizará un corpus pequeño para la generación de modelos y el entrenamiento, cuyas características se explicarán más adelante.
- El vocabulario será mediano, y se limitará a un conjunto de palabras con las que se podrán formar un número pequeño de frases ricas fonéticamente, para así poder probar el reconocimiento de habla continua.
- No se hará especial hincapié en la robustez frente a las condiciones adversas, por lo que será sensible al ruido externo.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

# 3 DESCRIPCION DEL SISTEMA



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

En este capítulo se presenta el desarrollo del Módulo de Reconocimiento de Habla Continua, indicando las técnicas matemáticas utilizadas en cada etapa, además del conjunto de herramientas (toolkit) que fue utilizado para dicha implementación.

## *3.1 HTK*

HTK (Hidden Markov Models Toolkit) es un conjunto de herramientas desarrolladas por la Universidad de Cambridge, Inglaterra, las cuales han sido diseñadas para construir módulos de procesamiento de habla basados en HMM (Hidden Markov Models, Modelos Ocultos de Markov), especialmente reconocedores, [8] cuyo diagrama general se muestra en la figura 3.1.

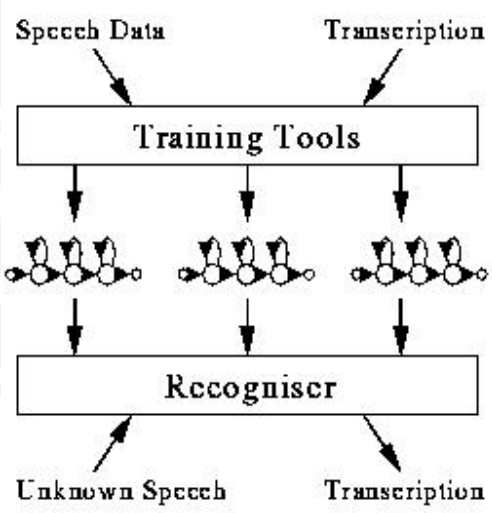



**Figura 3-1 Diagrama general de un reconocedor implementado con HTK [8]**



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

## 3.2 Pasos a seguir

Para crear un sistema de procesamiento de habla, ya sea de reconocimiento o síntesis, nos inspiramos en un proceso conocido, que es la comunicación hablada o diálogo entre dos personas. En la figura 3.2., observamos a dos personas dialogando, y representando alternadamente el papel de emisor y receptor.

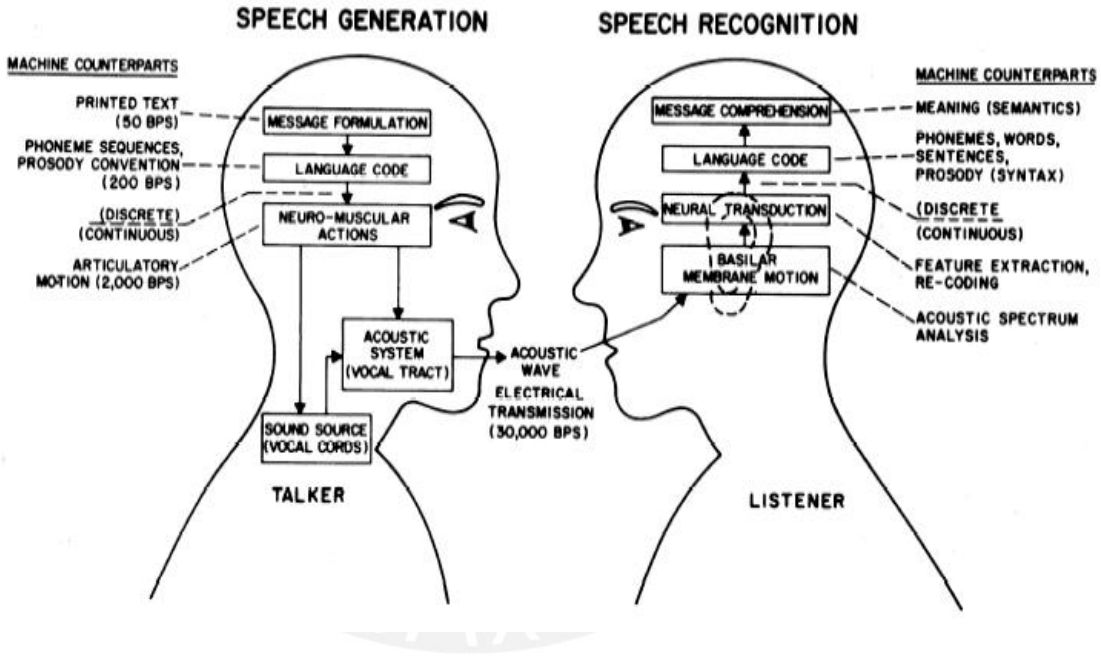



**Figura 3-2 Esquema del proceso de producción y percepción del habla (Rabiner y Juang, 1993)**

Se observa un paralelismo entre los procesos mentales del oyente y los pasos necesarios para implementar el reconocimiento de habla en una máquina, por lo que se concluye que las etapas de un sistema de reconocimiento de voz son :

- Análisis espectral de la señal de habla (extracción de características).



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Comparación de la señal con modelos de unidades acústicas ya establecidos (modelado de unidades, búsqueda de la hipótesis más probable y decodificación de la señal).
- Búsqueda de significado de la señal (comprensión del mensaje).

En el módulo a desarrollar, sólo se han considerado las dos primeras etapas, por considerarse que la tercera se enmarca mejor en el módulo de Comprensión de Lenguaje Natural. No obstante, hemos considerado el agregar una etapa previa al desarrollo propiamente dicho del módulo, la cual es la etapa de preparación de la base de datos que se utilizará en el entrenamiento y en las pruebas del sistema.

En conclusión, el módulo seguirá los siguientes pasos :

- Preparación de la base de datos o corpus de voces.
- Procesamiento de la señal de habla.
- Generación de modelos y entrenamiento.
- Decodificación de la señal recibida.

Esto puede apreciarse en la figura 3.3.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

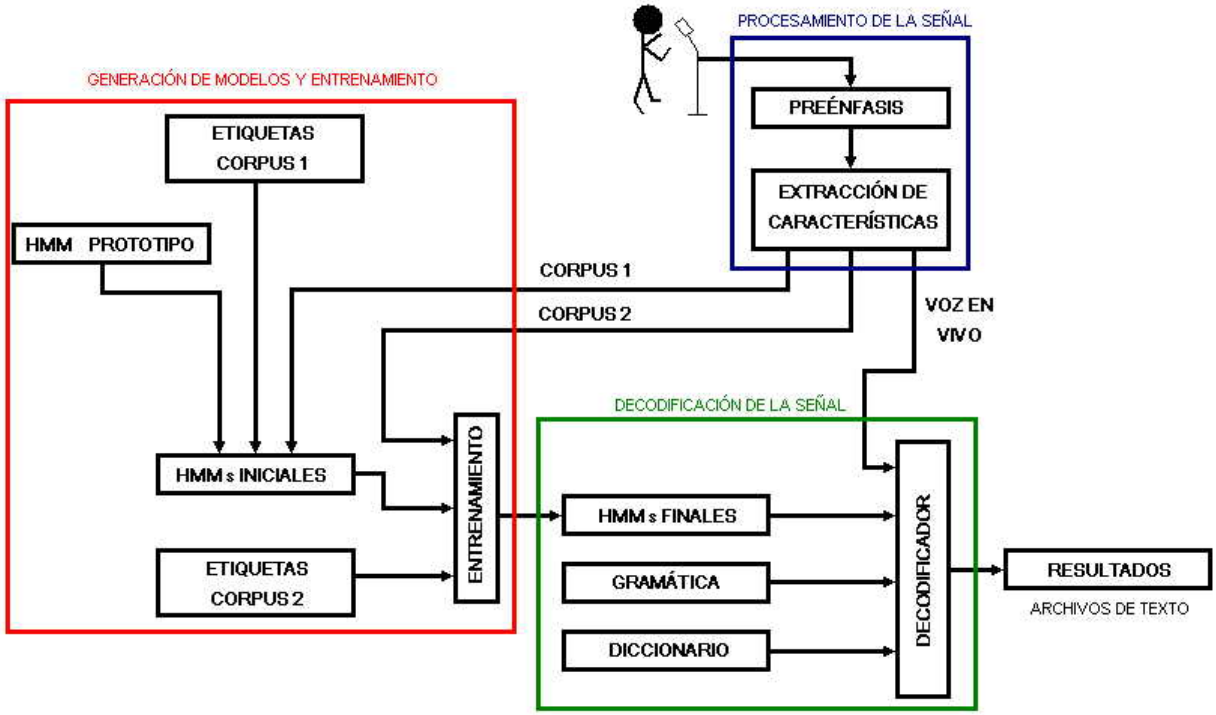



**Figura 3-3 Esquema del módulo de reconocimiento de habla continua**

### 3.2.1 Preparación de la base de datos

En el capítulo anterior se mencionó que el sistema debe ser independiente de locutor, lo cual implica que debe ser entrenado con la mayor cantidad y variedad posible de voces, de preferencia de todos los rangos de edad, estratos sociales y niveles de educación posibles. Es por ello que como paso previo, debe adquirirse o recolectarse un conjunto o **corpus** de voces en el idioma deseado, teniendo en cuenta :

- Las unidades que el sistema utilizará como patrones. Si se utilizan palabras, el corpus deberá contener varias muestras de cada palabra considerada; si se



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

utilizan unidades inferiores a la palabra, será la misma situación, sin embargo, el primer corpus será mucho mayor al segundo, ya que varias unidades pueden encontrarse en una sola palabra.

- El dominio en el que se desenvolverá el sistema. Por ejemplo, si va a dar información sobre viajes, debe incluir nombres de países, horas, fechas, etc. Si se destina al control de un aparato, bastarán palabras utilizadas como órdenes.

Actualmente existen varios corpus a disposición de los interesados en el área de tratamiento del habla. Algunos de estos son :

- AURORA, diseñado originalmente para establecer un estándar mundial para el software de extracción de características, que es el núcleo del front-end de todo sistema de reconocimiento distribuido de habla (Distributed Speech Recognition – DSR). Hay varias versiones de AURORA, tomadas en distintas condiciones e idiomas.

- English SpeechDat(M), que contiene grabaciones de 1000 hablantes, recogidas a través de una red de telefonía fija. Cada hablante da muestras de secuencias de números y letras, palabras de control, fechas, horas, cantidades de dinero, etc. Ha sido diseñado para entrenar y evaluar sistemas de reconocimiento independientes de hablante, ya sea de habla continua o de palabras aisladas.

- EUROMI (base de datos de habla europea y multilingue), la cual tiene corpus equivalentes para varios idiomas europeos (italiano, inglés británico, alemán, holandés, danés, sueco, noruego, francés, griego, español y portugués). Contiene más de 60 hablantes para cada idioma, los cuales fueron elegidos al azar.

- Albayzin (español), el cual consta de 3 sub-corpus, con un total de 304 hablantes del español de Castilla.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- SALA (SpeechDat Across Latin America), el cual recoge corpus de varias zonas de América Latina, cubriendo todas las regiones dialectales que representan las diferentes variaciones del español y portugués, para lo cual se dividió a América Latina en ocho zonas principales :
    - Brasil
    - México
    - Venezuela y el Caribe
    - América Central
    - Panamá y Colombia
    - Ecuador, Perú y Bolivia
    - Chile
    - Argentina, Uruguay y Paraguay

  Este proyecto aún no se ha completado, y la zona en la que se encuentra nuestro país aún no ha sido grabada. [10]

No obstante, los precios de estos corpus son un obstáculo para su uso, ya que oscilan entre los 250 y 12000 euros [9], además de que no son del español que se habla en el Perú, y más concretamente en Lima. Es por ello que se decidió generar un corpus propio.

#### 3.2.1.1 Selección de las unidades a utilizarse

Las unidades a utilizarse son aquellas que serán empleadas como patrones de comparación por el módulo. Para seleccionar cuál es la más apropiada, se tienen en cuenta tres factores [1]:



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- La unidad debe ser *precisa*, para poder representar la realización acústica que aparece en diferentes contextos.
- La unidad debe ser *entrenable*, es decir, se debe disponer de suficientes datos para estimar sus parámetros.
- La unidad debe ser *generalizable*, de tal manera que cada palabra nueva pueda ser derivada a partir de un inventario predeterminado de unidades.

Existen varias unidades, por ejemplo : fonos, fonemas, sílabas, demisílabas, palabras, frases…

Como el módulo debe ser independiente de locutor, es mejor utilizar unidades inferiores a la palabra, para poder generalizar el alfabeto a todos los posibles usuarios. Entre estas tenemos a los fonos[1], fonemas[2], sílabas y demisílabas. Los fonos y fonemas presentan una ventaja frente a las sílabas y demisílabas, que es mostrada en la tabla 3.1 :

[1] **Fono** : Los fonos son todos los sonidos posibles que un ser humano puede producir, los cuales son infinitos y han sido clasificados por la IPA ( Alfabeto Fonético Internacional ) [6]

[2] **Fonema y Alófono** : Un fonema puede definirse como una familia de sonidos, los cuales son considerados iguales por hablantes de una lengua específica. Los sonidos que forman parte de un mismo fonema se llaman alófonos o variantes alofónicas. Por ejemplo, en el idioma español, el fonema /b/ tiene dos alófonos : [b] y [B] . Si en una palabra donde esté presente [b], intercambiamos dicho alófono por [B] ( o viceversa), es decir, la pronunciamos mal, el oyente escuchará la misma palabra, porque ambos son el mismo fonema. Al contrario, si reemplazamos un fonema por otro ( /d/ en lugar de /b/, por ejemplo), el oyente escuchará otra palabra. [6]



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

| Unidad | Número de elementos |
|---|---|
| Fonos, fonemas | 30 a 50 |
| Sílabas | 8000 a 10 000 |
| Demisílabas | 1000 a 2000 |

**Tabla 3-1 Unidades versus Número de elementos**

Como se observa , el número de fonos es mucho menor , es decir se tiene que modelar o entrenar menor número de patrones o modelos que en el caso de sílabas o demisílabas, lo que implica menor cantidad de material de entrenamiento y menor tiempo de estimación de los modelos. Sin embargo, es necesario notar que no son muy precisas, ya que se asume que los fonos no son afectados por los fonos que lo preceden y que lo siguen, es decir, por el contexto.

Actualmente, se suelen usar alófonos llamados *trifonos*, los cuales son modelos fonéticos que consideran la influencia de los fonos situados a la izquierda y a la derecha. Si dos fonos tienen el mismo nombre, pero diferentes vecinos a la derecha y a la izquierda, entonces son considerados trifonos diferentes.

El utilizar trifonos mejora la tasa de reconocimiento del sistema, ya que modelan mejor los efectos coarticulatorios [1]. No obstante, dado su mayor número, se necesita una mayor cantidad de datos para entrenarlos, lo cual nos da otra razón para utilizar sólo fonos separados, sin tomar en cuenta el contexto.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

### 3.2.1.2 Construcción del alfabeto fonético

Una vez escogidas las unidades, se procede a armar el alfabeto con el que se realizan las transcripciones fonéticas de las emisiones de habla que se destinarán al entrenamiento.

Para el módulo, se ha elegido un alfabeto fonémico, el cual consta de 29 unidades, en el cual se han agrupado varios alófonos bajo un mismo nombre, para minimizar en lo posible la carga de procesamiento del sistema. El alfabeto es presentado en la tabla 3.2.

| Clase | Descripción | SONORO | SORDO |
|---|---|---|---|
| OCLUSIVO | Explosión producida en un punto donde los órganos articuladores hacen contacto. | b bino | p padre |
| | | d donde | t tomo |
| | | g gata | k kasa |
| FRICATIVO | Alófonos de oclusivo ( B, D,G ):Los puntos de oclusión ceden hacia una leve abertura donde el paso del aire produce una ligera fricción //Fricativas y sibilantes: el aire al pasar por una breve abertura hace una fuerte fricción (turbulencia) | bb kabbra | f fasil |
| | | dd nadda | s sine |
| | | gg lueggo | x muxer |
| | | | ss assaninka |
| NASALES | El velo del paladar baja permitiendo el paso del aire por la cavidad nasal. La cavidad bucal sirve para la resonancia, | m amos | |
| | | n uno | |
| | | J niJo | |
| | | nn kannxe | |
| LIQUIDAS | Vibrantes: oclusión bucal momentánea o de rápida intermitencia. //Laterales: oclusión bucal apico dental, el aire sale | r arma | |
| | | rr rrosa | |



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

| | | | |
|---|---|---|---|
| | por uno o los dos lados. | l ala | |
| SEMI CONSONANTE | Una vocal se desplaza de una abertura a un cierre (semiconsonante) y viceversa. | y yanta<br>w awto | |
| VOCALES | No hay obstrucción de ningún tipo.<br>La lengua sube o baja modelando la vocal. | a casa<br>e texa<br>i piko<br>o torre<br>u luna | |
| COMPUESTAS | Son dos fonemas que el hablante percibe como unidad. | xx exxito<br>C CanCo | |

**Tabla 3-2 Alfabeto fonémico**

### 3.2.1.3 Dominio del sistema

Como este módulo estará enfocado a reconocer un pequeño número de frases con la mayor precisión posible, las frases que serán recolectadas serán generales y construidas de tal manera que todas las unidades que componen el alfabeto estén presentes.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

### 3.2.1.4 Diseño del corpus de voces

#### *3.2.1.4.1 Características*

Se buscó que el corpus fuera tomado en un ambiente lo más libre de ruido que se pudiera, para poder obtener modelos tan desprovistos de alteraciones que no fueran provenientes del propio hablante y del transductor utilizado. Para ello, se realizaron sesiones de grabación en el estudio de radio de la Facultad de Ciencias y Artes de la Comunicación, de la Pontificia Universidad Católica del Perú.

El transductor escogido fue un micrófono dinámico AKG, modelo D 3800, que tiene las siguientes características :

- Patrón de radiación cardioide, lo cual bloquea en parte el ruido indeseado, al no captar el sonido que se produce detrás del micrófono (Figura 3.4. (a)).
- Respuesta de frecuencia bastante plana en el intervalo de 300 a 3000 Hz, lo que lo hace ideal para aplicaciones que tengan que ver con voz (Figura 3.4. (b)).

Además del micrófono, se utilizaron un cable blindado con conector XLR, una consola de audio MONA, fabricada por Echo Digital Audio Corp, el software CoolEdit 2000, ahora Adobe Audition, y una computadora Pentium IV, con 256 MB de memoria.

El corpus en sí es una lista de 260 palabras, escogidas de tal manera que estuvieran presentes todos los fonos del idioma español hablado en Lima, Perú, en todas las posiciones posibles (inicio, mitad o final de palabra), lo cual hace un total de aproximadamente 5 minutos de grabación por persona.

Los parámetros de grabación fueron los siguientes :



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Formato de sonido : WAV
- Tasa de muestreo : 8000 Hz
- Número de canales : 1 (Mono)
- Bits por muestra : 16

Gracias a dichos parámetros, obtenemos muestras con un alto contenido espectral dentro de los límites de la voz (entre 2 Hz y 20 KHz), suficientemente grandes para poder obtener la informacion suficiente para el análisis pero lo suficientemente pequeñas como para no ocupar mucho espacio de almacenamiento.

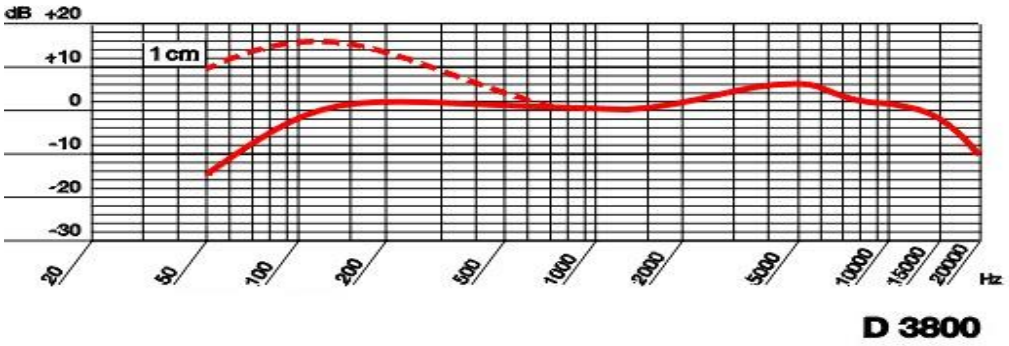
En total, se tomaron grabaciones de un total de 150 hablantes, lo cual le dio al corpus las siguientes características :

- Proporción de hombres/mujeres : 2/1
- Rango de edad predominante : 19 a 23 años
- Nivel de educación : la mayoría pertenece a la Pontificia Universidad Católica del Perú, en el nivel de pregrado.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

(a)




(b)

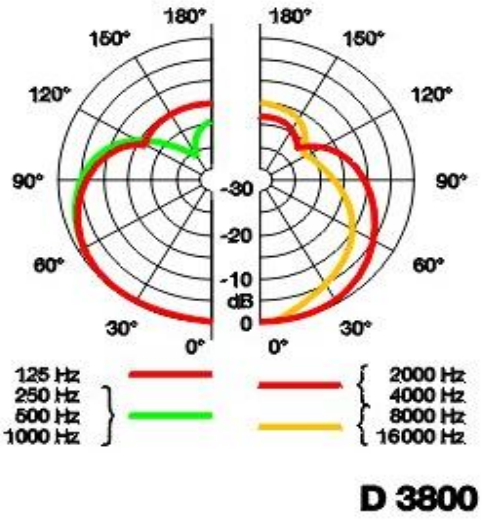



**Figura 3-4 Características del micrófono AKG D 3800**

**(a) Diagrama de frecuencias, (b) Patrón de radiación**



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

#### *3.2.1.4.2 Segmentado y etiquetado*

Una vez recolectado el corpus, se procedió a segmentarlo y etiquetarlo, lo cual consiste en dividir cada palabra en sonidos y generar un archivo de texto que contenga etiquetas para cada una de estas divisiones.

Estas etiquetas siguen un formato especial, llamado TIMIT, el cual es el formato utilizado en la base de datos TI-MIT, creada en 1989 por Texas Instruments y el Instituto Tecnológico de Massachussets (MIT). En TIMIT, cada línea del archivo de texto contiene la siguiente información :

| | | |
|---|---|---|
| <número de muestra del comienzo de la palabra o sonido> | <número de muestra del final de la palabra o sonido> | <nombre de la palabra o sonido> |

Podemos ver un ejemplo de esto en la figura 3.5.

```
15960 17200 SIL
17200 18960 b
18960 20880 i
20880 22720 n
22720 25400 o
25400 29720 SIL
```

**Figura 3-5 Archivo TIMIT para la palabra “vino”**

La segmentación fue realizada de manera manual, por el señor Giancarlo Peña, alumno de la especialidad de Lingüística, de la Pontificia Universidad Católica del Perú. Si bien este tipo de segmentación es más exacto, tiene el inconveniente de que toma demasiado tiempo y es una tarea tediosa.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

## 3.3 Procesamiento de la señal de habla

Esta etapa consiste en tratar la señal de habla que se recibe del transductor, una vez pasada por un conversor análogo-digital, con una serie de técnicas matemáticas, para extraer su información espectral y comprimirla en un vector, sin sacrificar información fonética o del hablante y manteniendo el orden temporal.

Esta etapa se divide en dos procesos, preénfasis y extracción de características, tal como muestra la figura 3.6.

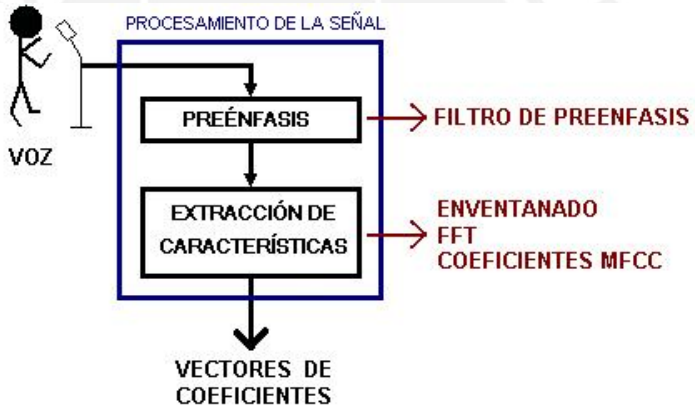



**Figura 3-6 Etapa de procesamiento de la señal**

### 3.3.1 Preénfasis

El proceso de preénfasis consiste de un filtro pasa-alto de primer orden cuyo propósito es compensar el efecto del aire que atenúa las frecuencias altas.

El filtro de preénfasis tiene la siguiente fórmula :



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

$$H(z) = 1 - az^{-1}, \quad 0.9 \le a \le 1.0$$

Un valor típico para “a” es 0.97.

### 3.3.2 Extracción de características

La extracción de características consiste en extraer las características espectrales de la señal, para generar los llamados “vectores de coeficientes”, los cuales son conjuntos de coeficientes calculados utilizando técnicas de predicción lineal.

Para ello, los pasos son :

1. Enventanamiento de la señal [2]

El enventanamiento es la multiplicación de la señal de habla por una función ventana, lo cual da como resultado un conjunto de muestras alteradas por la forma de dicha ventana.

Para escoger una ventana apropiada, se debe tener en cuenta que debe ser lo suficientemente estrecha para que las propiedades de la señal cambien muy poco, pero a su vez lo suficientemente grande para que existan las muestras suficientes para calcular los parámetros deseados. Además, hay que tener en cuenta el número total de ventanas, el cual debe ser lo suficientemente grande como para no perder información de la señal. Esta última condición se refleja más en la *tasa de ventanas* (número de veces por segundo en que se realiza el análisis de la señal, mientras se hace avanzar la señal en el tiempo de manera periódica), la cual es normalmente el doble de la inversa de la duración de la ventana.

La mayoría de las aplicaciones utilizan ventanas grandes, que enfatizan la información que cae en el medio de ellas. Por ejemplo, si el habla es cuasi-estacionaria alrededor de



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

los 10 ms, una ventana de 20 ms dejará pasar los 10 ms centrales con menos distorsión que los 5 ms del inicio y del final.

La alternativa más común dentro de las ventanas es la ventana Hamming, cuya ecuación es :

Para valores fuera del rango de 0 a n-1, la amplitud es igual a cero.

Para el módulo, se ha escogido la ventana de Hamming, con una longitud de 20 ms.

2. Cálculo de los vectores de coeficientes

La última etapa del procesamiento de la señal de voz es la extracción de vectores de características o parámetros. Para ello se dispone de varias técnicas, las cuales se dividen en dos grupos :

- Parámetros derivados del espectro de Fourier (MFCC, LFCC)
- Parámetros derivados del espectro LPC (LPCC, LPC, RC)

Los parámetros derivados del espectro de Fourier preservan información de la señal que es omitida por el segundo grupo. Entre aquellos del primer grupo, los coeficientes Mel-Frequency cepstrum (MFCC) son los más empleados en las aplicaciones de reconocimiento de habla, ya que poseen mayor robustez frente al ruido y a las variaciones de estimación espectral. [1]



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

Los MFCC se definen como la parte real del cepstrum de un segmento de señal, el cual ha sido enventanado y tratado con la DFT. Para calcularlos, la señal de voz atraviesa un banco de M filtros triangulares, donde la suma de sus anchos de banda cubre el espectro de la voz, de manera no lineal. Esto es debido al comportamiento que presenta el sistema auditivo humano; según estudios, la cóclea se comporta como un conjunto de filtros traslapados entre sí, los cuales se distribuyen a lo largo de bandas críticas de manera no lineal.

La función de transferencia de estos filtros se calcula en la siguiente ecuación :

$$H_m[k] = \begin{cases} 0 & , k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & , f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & , f[m] \le k \le f[m+1] \\ 0 & , k > f[m+1] \end{cases}$$

donde k proviene de la fórmula de la DFT, y f[m] se calcula con la siguiente ecuación :

$$f[m] = \left(\frac{N}{fs}\right) B^{-1}\left(B(f_L) + m\frac{B(f_H) - B(f_L)}{M+1}\right)$$

donde $f_L$ y $f_H$ son las frecuencias máxima y mínima del ancho de banda total de la voz, expresadas en Hz, fs es la frecuencia de muestreo, M el número total de filtros y N el tamaño de la DFT. Podemos observar que $f_L$ y $f_H$ están escaladas por B, que representa a la escala mel, y $B^{-1}$, que es su inversa.



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

La escala mel es una escala de frecuencias no lineal, y se caracteriza por tener un comportamiento casi lineal por debajo de 1 KHz y logarítmico por encima. Los valores de frecuencias en esta escala se encuentran aplicando la ecuación:

$$B(f) = 1125 Ln\left(1 + \frac{f}{700}\right)$$ con f en Hz.

Podemos ver el banco de filtros en la figura 3.7.

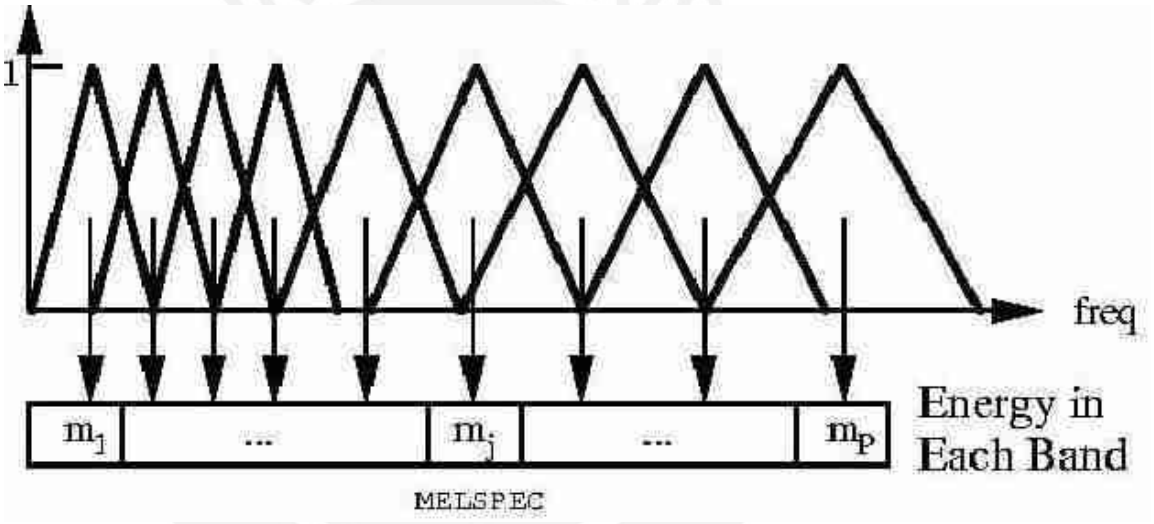



**Figura 3-7 Banco de filtros triangulares utilizados en el cálculo de los coeficientes MFCC [8]**

Luego de hacer pasar la señal por este banco de filtros, calculamos el logaritmo de la energía a la salida, con la siguiente ecuación :

$$S[m] = Ln\left(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m(k)\right), \quad 0 < m \leq M$$



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

Finalmente, aplicando la Transformada Coseno Discreta (DCT), obtenemos los coeficientes deseados :

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left( \frac{\pi n \left( m - \frac{1}{2} \right)}{M} \right), \; 0 \le n < M$$

Generalmente, para un sistema de reconocimiento se toman los primeros 13 coeficientes, a los cuales se agregan sus derivadas en el tiempo, para considerar los cambios temporales en el espectro de la señal. Estas derivadas se llaman coeficientes delta (primera derivada) y aceleración (segunda derivada).

### 3.3.3 Implementación en HTK

Para esta primera etapa, empleamos la herramienta HCopy [8], especificando los valores deseados por medio de un archivo de configuración, como el que se presenta en la figura 3.8. Podemos observar que la señal de entrada es de tipo WAV, muestreada a 8KHz (expresado en unidades de 100 ns), y la salida será 12 coeficientes MFCC, a los cuales se agregarán el coeficiente $C_0$ (logaritmo de la energía), además de los coeficientes delta y aceleración, lo que hace un total de 39 coeficientes.

Se utilizará una ventana Hamming, de 25 ms (en unidades de 100 ns), un filtro de preénfasis con coeficiente igual a 0.97 y un banco de 26 filtros triangulares.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

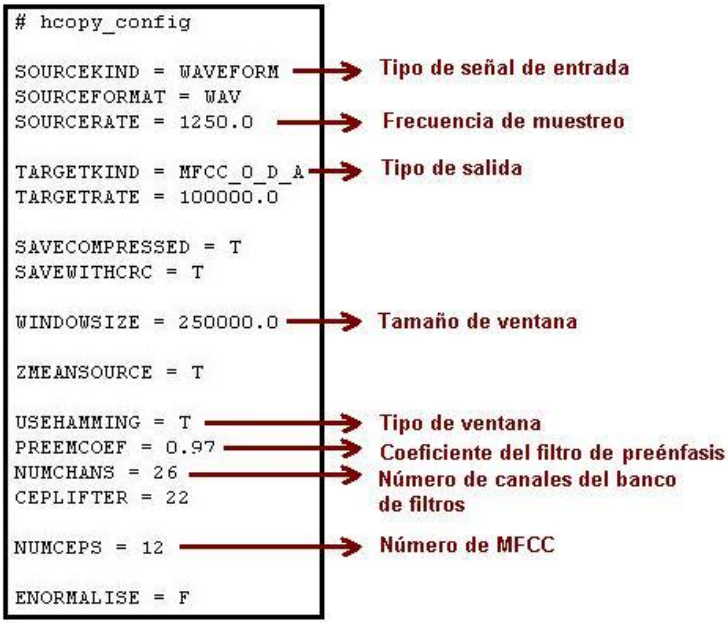



**Figura 3-8 Archivo de configuración para HCopy**

Los vectores correspondientes a cada señal son almacenados en el directorio “mfcs” en archivos con extensión “.mfc”.

Para ver los resultados, se emplea la herramienta HList [8], tal como se ve en la figura 3.9.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
------------------------------ Source: timit.wav -------------------------
  Sample Bytes:  2        Sample Kind:   WAVEFORM
  Num Comps:     1        Sample Period: 62.5 us
  Num Samples:   31437    File Format:   TIMIT
------------------------------------ Target ------------------------------
  Sample Bytes:  72       Sample Kind:   MFCC_E_D
  Num Comps:     18       Sample Period: 10000.0 us
  Num Samples:   195      File Format:   HTK
-------------------------- Observation Structure --------------------------
x:     MFCC-1   MFCC-2   MFCC-3   MFCC-4   MFCC-5   MFCC-6   MFCC-7   MFCC-8        E
        Del-1    Del-2    Del-3    Del-4    Del-5    Del-6    Del-7    Del-8     DelE
------------------------------ Samples: 100->104 ------------------------
100:    3.573 -19.729  -1.256  -6.646  -8.293 -15.601 -23.404  10.988   0.834
        3.161  -1.913   0.573  -0.069  -4.935   2.309  -5.336   2.460   0.080
101:    3.372 -16.278  -4.683  -3.600 -11.030  -8.481 -21.210  10.472   0.777
        0.608  -1.850  -0.903  -0.665  -2.603  -0.194  -2.331   2.180   0.069
------------------------------------ END ---------------------------------
```

**Figura 3-9 MFCC vistos con HList**

## *3.4 Generación de modelos y entrenamiento*

En esta etapa se generan los modelos o patrones de cada una de las unidades que conforman el alfabeto elegido.

Para generar los modelos iniciales se ha elegido una técnica estadística, denominada Modelos Ocultos de Markov (Hidden Markov Models – HMM), la cual presume que las muestras de habla pueden ser caracterizadas por un proceso paramétrico y aleatorio, y que dichos parámetros pueden ser estimados en un marco preciso y bien definido.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

### 3.4.1 Modelos Ocultos de Markov (HMM) [5] [1]

Un Modelo Oculto de Markov (Figura 3.10) es una máquina de estados finitos, con entradas conocidas, salidas como una función probabilística de las entradas y una secuencia de estados desconocida u "oculta". Esto quiere decir que la máquina de estados en sí es una caja negra; no conocemos qué secuencia de estados genera una determinada salida (generador de vectores).

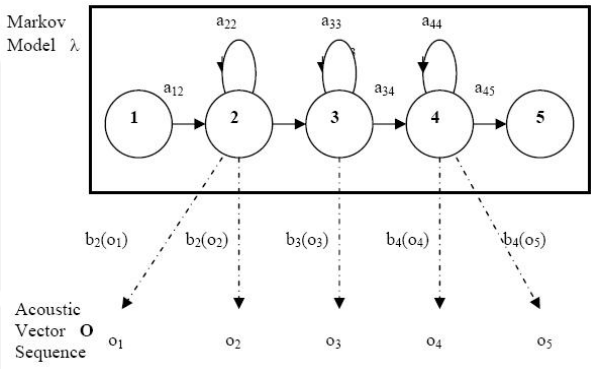



**Figura 3-10 HMM de 5 estados**

Un HMM está definido por:

- Ω={1, 2, ..., N}, conjunto de estados del modelo, donde $q_t$ representa el estado en el tiempo *t*.
- O={$o_1$, $o_2$, ..., $o_M$}, número de símbolos observados por estado, que corresponden a la salida física del sistema.
- A, matriz de distribución de probabilidades de transición entre estados

$$a_{ij} = P[q_{t+1} = j / q_t = i], \quad 1 \le i, j \le N$$



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

- B, matriz de distribución de probabilidades de obtener un determinado símbolo a la salida

$$b_j(k) = P[o_t = v_k / q_t = j] \quad 1 \le k \le M$$

- $\pi$, matriz de distribución inicial de estados

$$\pi_i = P[q_1 = i], \quad 1 \le i \le N$$

Desde que $a_{ij}$, $b_j(k)$ y $\pi_i$ son probabilidades, deben satisfacer las siguientes propiedades :

$$a_{ij} \ge 0, \; b_i(k) \ge 0, \; \pi_i \ge 0 \quad \forall i, j, k$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

$$\sum_{k=1}^{M} b_i(k) = 1$$

$$\sum_{i=1}^{N} \pi_i = 1$$

Para representar estos parámetros, se utiliza la siguiente notación :

$$\Phi = (A, B, \pi)$$



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

### 3.4.1.1 Problemas al construir un HMM [5]

Los HMMs presentan tres problemas principales :

- Evaluación de probabilidades, lo cual implica el cálculo de la probabilidad de que una salida, también llamada “observación”, se obtenga con un modelo λ. La evaluación se realiza con el algoritmo Forward-Backward.
- Secuencia “óptima” de estados, cuya solución es la implementación del algoritmo de Viterbi, ya que este nos permite encontrar el mejor camino posible, evaluando la probabilidad.
- Estimación de parámetros, de manera óptima. El algoritmo de Baum-Welch, también llamado EM (expectation-maximization), nos permite optimizar los parámetros necesarios para construir el modelo.

### 3.4.1.2 Topologías de HMM [5]

Los HMMs se clasifican, de acuerdo a la estructura de su matriz de transición, en :

- Ergódico, en el cual un estado puede ser alcanzado desde otro estado en un número finito y aperiódico de saltos.
- Izquierda-derecha o de Bakis, cuya característica es que los saltos siempre van únicamente de izquierda a derecha, y como máximo puede alcanzar a un estado distante dos saltos.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

El sistema utiliza HMM tipo izquierda-derecha, como el de la figura 3.11., la cual es la más utilizada en aplicaciones de reconocimiento de habla [1].

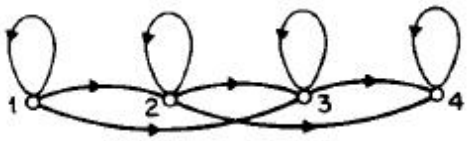



**Figura 3-11 HMM izquierda-derecha (left-to-right)**

Para este tipo de topología, lo más importante es el número de estados que se van a utilizar, ya que depende de dos factores [1] :

- La cantidad de datos de la que se disponga para el entrenamiento.
- El uso que se le va a dar al modelo; por ejemplo, si se desea representar un fonema, se debe tener como mínimo tres o cinco estados. Si se desea representar una palabra, debe tenerse en cuenta la pronunciación y duración de ella para determinar el número de estados. Una excepción es el silencio, ya que debido a su naturaleza estacionaria, basta un HMM de uno o dos estados para modelarlo.

### 3.4.2 Entrenamiento

El entrenamiento consiste en la reestimación de los parámetros de los HMM realizando varias iteraciones, empleando un corpus de voces debidamente segmentado y etiquetado.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

### 3.4.3 Implementación en HTK

Para la generación de modelos, hemos empleado la herramienta HInit [8], que requiere de un archivo "prototipo" (Figura 3.12), el cual será la plantilla para los HMMs iniciales. En este archivo se define el número de estados, la media y la varianza, sin importar los valores.

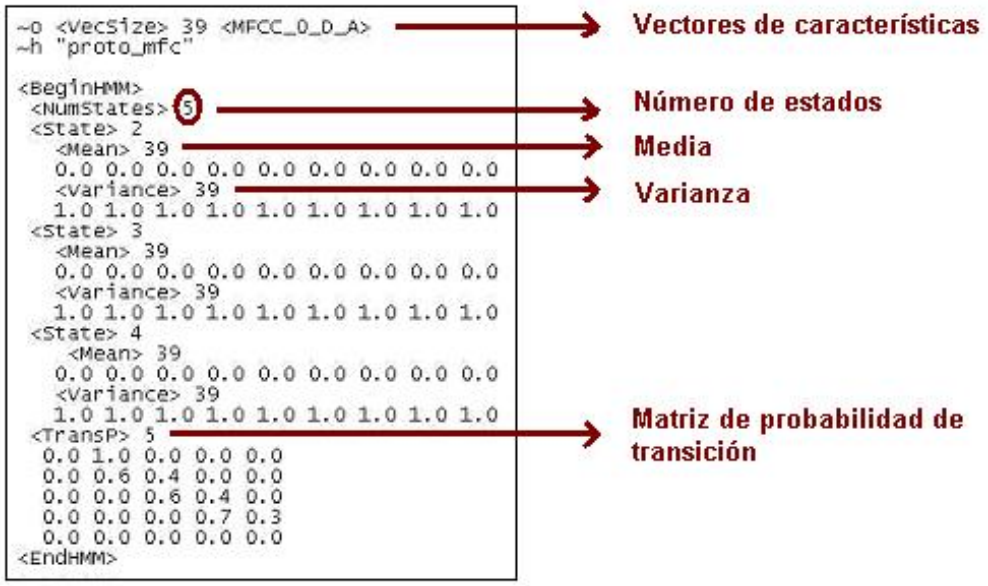



**Figura 3-12 Archivo prototipo para un HMM**

Hemos elegido un HMM de 5 estados, usando vectores de 39 elementos y una matriz de probabilidad de transición de 5 x 5.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

Junto con el corpus destinado a la generación de modelos (señales WAV y archivos TIMIT), se crean los HMM para cada uno de los fonemas del alfabeto.

El entrenamiento se realiza con las herramientas HRest [8] y HERest [8], las cuales realizan el entrenamiento individual y global respectivamente.

El proceso seguido se observa en la Figura 3.13.

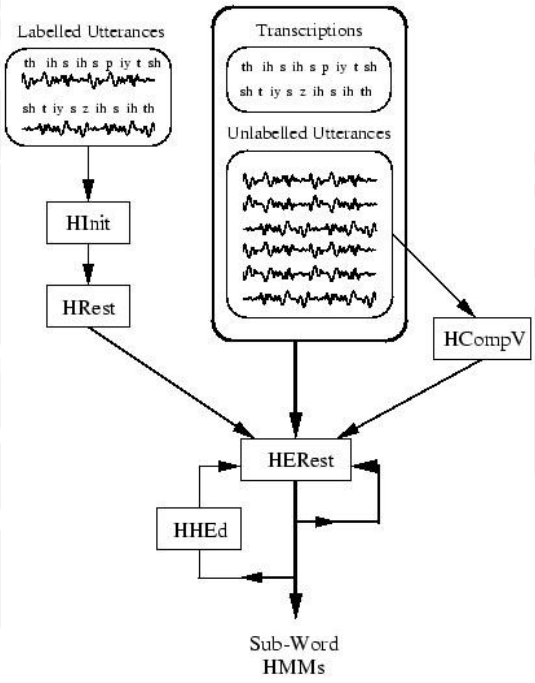



**Figura 3-13 Diagrama de generación y entrenamiento**

Al final, obtenemos los HMM finales, almacenados en el directorio “hmms3”.



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

## 3.5 Decodificación de la señal

Esta etapa es el reconocedor propiamente dicho, el cual tiene como entradas los HMM finales o patrones de comparación; la gramática, que es un conjunto de reglas de construcción; el diccionario [1], donde se almacenan la mayor variedad posible de pronunciaciones por palabra, y la voz a analizar, la cual puede ser un archivo WAV o una voz "en vivo".

El decodificador se basa en el algoritmo de Viterbi [1], que también emplea probabilidades y busca el modelo o la combinación de ellos que se ajusten mejor a la señal ingresada.

El esquema de esta etapa se observa en la figura 3.14.

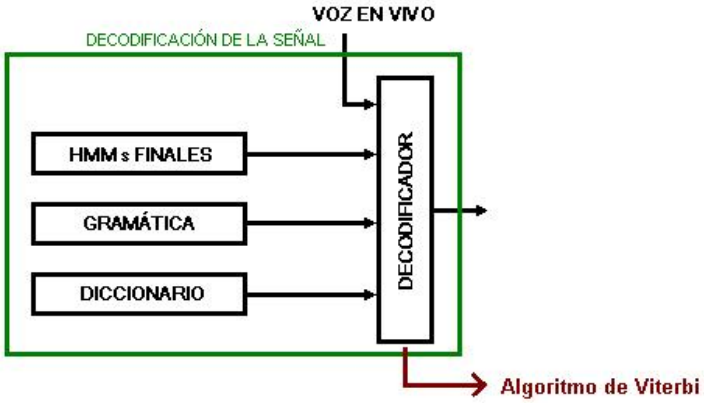



**Figura 3-14 Decodificador**



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

### 3.5.1 Implementación en HTK

La implementación consiste en crear las entradas que faltan (gramática y diccionario) y de implementar un algoritmo que permitirá escoger la combinación que tenga el parecido más próximo a la señal de audio analizada.

#### 3.5.1.1 Creación de la gramática

Para crear la gramática, utilizamos la herramienta HParse [8], junto con un archivo donde se incluyen las palabras del vocabulario a reconocer, separadas por clases, y las reglas gramaticales que ha de seguir. La salida es una red de palabras, donde se reflejan todas las posibles relaciones entre las palabras del vocabulario.

Para el módulo, se ha optado por la gramática más sencilla: todas las palabras pertenecen a una única clase, con las que pueden armarse varias combinaciones, con un silencio largo al inicio de la emisión y otro al final.

#### 3.5.1.2 Creación del diccionario

El diccionario es un archivo de texto donde se incluyen las palabras del vocabulario y su transcripción fonética, utilizando el alfabeto creado. Si existen variaciones de una misma palabra, también se incluyen. Por ejemplo, la palabra “adriana” figura de la siguiente manera :

| | | |
|---|---|---|
| Adriana | [adriana] | a d r y a n a |
| Adriana | [adriana] | a d r y a n a silf |
| Adriana | [adriana] | d r y a n a |
| Adriana | [adriana] | d r y a n a silf |
| Adriana | [adriana] | r y a n a silf |
| Adriana | [adriana] | r y a n a |

Este archivo de texto fue creado de manera manual, en colaboración con un estudiante de la especialidad de Lingüística.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

### 3.5.1.3 Decodificador

Finalmente, con la herramienta HVite [8] implementamos el algoritmo de Viterbi, que es el decodificador propiamente dicho, dando como entradas la lista de HMM, el diccionario, la gramática, la red de palabras y la señal de audio a analizar.

El resultado final es un archivo de texto (Figura 3.15), el cual muestra cuáles palabras se han pronunciado, cuáles modelos han sido detectados y el puntaje respectivo.

```
"./testdata/dif.rec"
0 1000000 SIL -664.797119 SIL_INICIO -75.000000
1000000 1300000 k -204.026306 camila -75.000000
1300000 2500000 a -1023.327637
2500000 2800000 m -259.148926
2800000 4000000 i -972.987732
4000000 4800000 l -568.101990
4800000 6300000 a -1155.477173
6300000 7800000 SIL -1133.728638
7800000 9600000 e -1085.439941 escribe -75.000000
9600000 11300000 s -1261.155151
11300000 12000000 sil -597.879639
12000000 12600000 k -452.427368
12600000 12900000 r -286.092133
12900000 14700000 i -1305.137329
14700000 15000000 bb -232.256073
15000000 16400000 e -889.101013
16400000 18800000 SIL -1806.652588
18800000 21500000 a -1859.855957 a -75.000000
21500000 24500000 SIL -2253.651367
24500000 24800000 m -364.194305 maquina -75.000000
24800000 26300000 a -1379.604736
26300000 26600000 k -272.393707
26600000 27700000 i -941.414978
27700000 28200000 n -397.515106
28200000 29200000 a -714.509644
29200000 31100000 SIL -1403.917358
```

**Figura 3-15 Salida del decodificador**



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

# 4 PRUEBAS Y RESULTADOS



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

El módulo se pondrá a prueba utilizando un conjunto de 83 voces, tanto de varones como de mujeres, de edades que varían entre los 20 y los 34 años, tomadas en el ambiente del laboratorio de Procesamiento Digital de Señales, ubicado en el tercer piso del pabellón V, Pontificia Universidad Católica del Perú.

Dicho ambiente tiene un nivel de ruido que va de los 20 a los 25 dB, y está compuesto de ruido propio de las computadoras (monitores, fuentes, discos duros), conversaciones en voz baja y música de fondo. Para evitar en lo posible la influencia de este ruido, el micrófono utilizado se colocó lo más lejos posible; no obstante, el ruido de la fuente de alimentación de la computadora se filtró a través de la clavija del micrófono. Debido a ello, se hizo una limpieza previa de cada emisión captada.

Cada emisión consta de 24 oraciones, leídas a velocidad normal, sin insistir en el cuidado al pronunciar ni en diferenciar claramente una palabra de otra, lo cual permite simular de manera muy aproximada el entorno real donde se debe desenvolver el módulo creado.

Una característica del corpus de prueba es que la proporción de varones y mujeres es muy parecida a la del corpus de entrenamiento; es decir, de 2 a 1, tal como se puede ver en la figura 4.1.

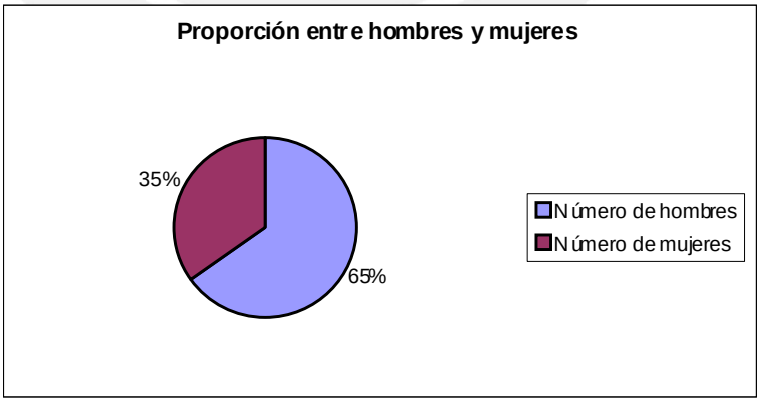



**Figura 4-1 Proporción entre hombres y mujeres en el corpus de prueba**



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

Para obtener los resultados en porcentajes, se utilizó la herramienta HResults [8], para lo cual se necesita que el corpus esté etiquetado. Es por ello que se segmentó y etiquetó manualmente sólo parte del corpus, debido a la gran cantidad de tiempo que requiere dicha operación. En total, se etiquetaron 164 frases, a dos frases por cada hablante.

HResults calcula tres tipos de errores al reconocer los fonos :

- Errores de sustitución (S), donde un fono es sustituido por otro.
- Errores de supresión (D), donde un fono es ignorado.
- Errores de inserción (I), donde se insertan fonos de más.

Con estos 3 valores, es posible calcular los siguientes porcentajes [8] :

- Porcentaje de corrección, que ignora los errores de inserción :

$$\%Corr = \frac{N - D - S}{N} \times 100\%$$

  donde N es el número total de etiquetas en las transcripciones de referencia.

- Porcentaje de precisión, el cual considera los 3 tipos de errores :

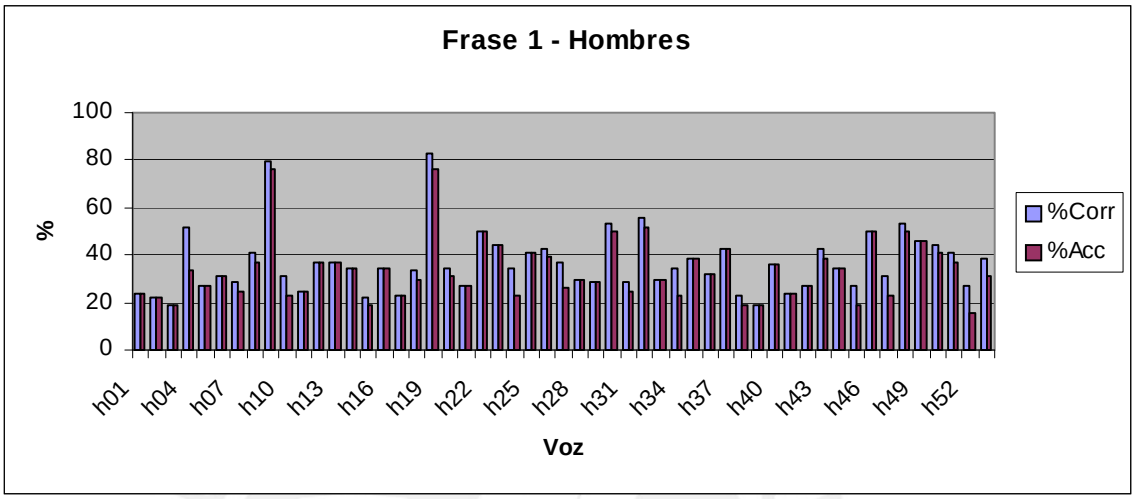
$$\%Acc = \frac{N - D - S - I}{N} \times 100\%$$

Las figuras 4.2 y 4.3 muestran los resultados obtenidos para 2 frases de prueba.



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP



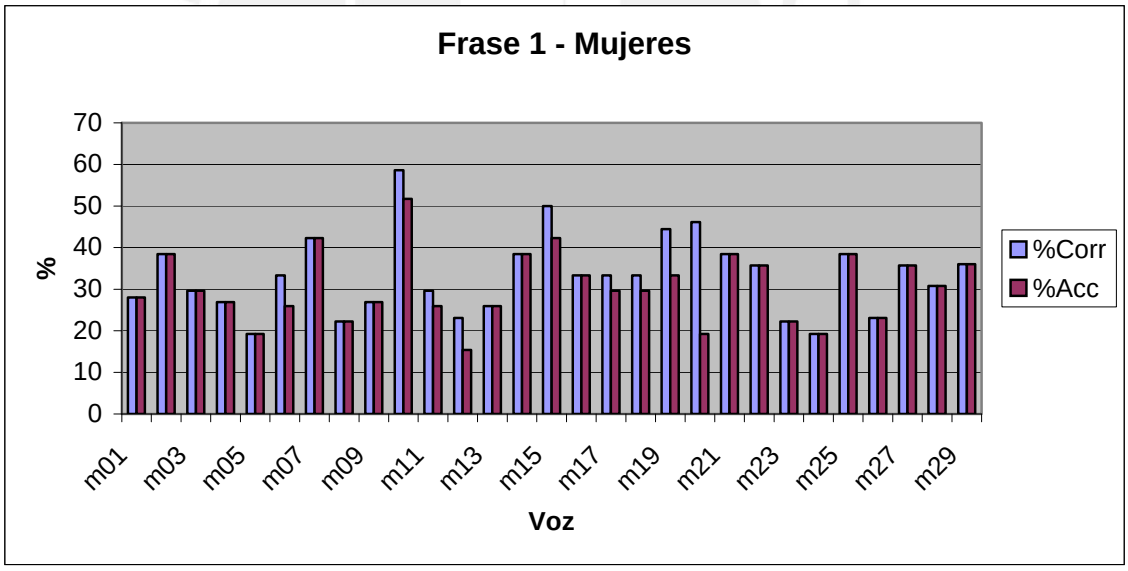
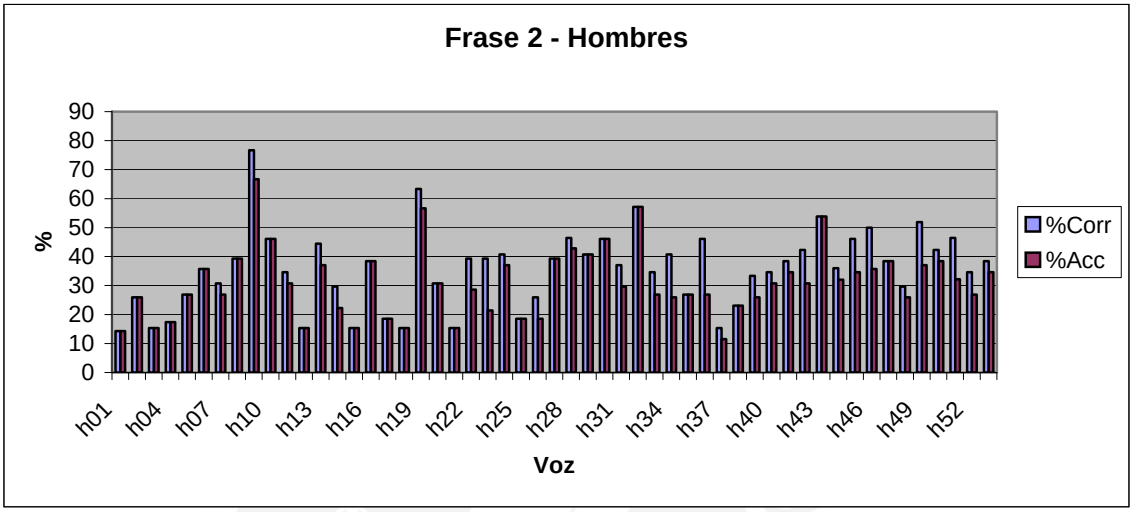




**Figura 4-2 Porcentajes de precisión y de corrección en hombres y mujeres (frase 1)**



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP



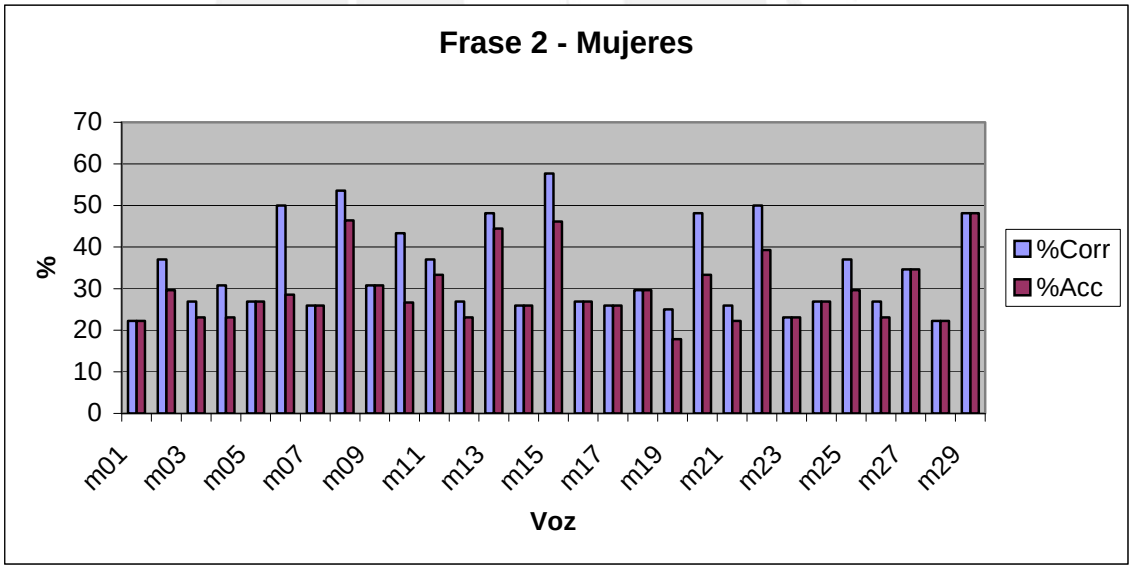




**Figura 4-3 Porcentajes de precisión y de corrección en hombres y mujeres (frase 2)**



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

En las pruebas realizadas utilizando una sola voz, se ha obtenido un 92% de precisión y un 86% de corrección, lo cual quiere decir que se han reconocido el 92% de los fonemas emitidos, pero sólo se obtienen un 86% de palabras correctas. Esto es debido a que la mayor parte de archivos usados para el entrenamiento correspondían a dicha voz en particular, debido a que ya estaban etiquetados.

En la pruebas realizadas utilizando 3 voces distintas, el porcentaje de precisión en el reconocimiento ha caído drásticamente, a un 20% de precisión. Debido a esto, la entrada que detecte el sistema no sería correcta la mayor parte del tiempo. Sin embargo, se ha notado que la tasa de aciertos mejora al incorporar al entrenamiento archivos nuevos, los cuales proceden del corpus más reciente. No obstante, el etiquetado consume tiempo, ya que se realiza manualmente.



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

# 5 OBSERVACIONES Y CONCLUSIONES



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Los mayores porcentajes pertenecen a los varones, debido a las características de su voz, la cual es más rica en frecuencias bajas. Estas frecuencias bajas se preservan mejor a la hora de digitalizar la voz por medio del micrófono, lo que permite la mejor conservación de la información espectral.

- En el caso de las mujeres, ya que la información de las frecuencias altas se altera al captar la voz, y las voces femeninas concentran la mayor parte de su información espectral en dichas frecuencias altas, su porcentaje de reconocimiento siempre será menor al de los varones.

- El diccionario de pronunciación tiene un papel muy importante en el proceso de reconocimiento, siendo necesaria una cuidadosa preparación de sus componentes, teniendo en cuenta que cada palabra puede tener más de una manera de pronunciarse, de acuerdo al español hablado en Lima.

- La gramática del sistema también juega un papel de importancia, ya que le da al sistema una plantilla por la cual guiarse para identificar una frase correctamente construida.

- Se observa que los hablantes tienden a suprimir las últimas vocales de cada palabra, así como a confundir fonos, lo cual hace que el sistema confunda una palabra con otra con mayor facilidad.

- Las unidades escogidas para el módulo no son muy adecuadas, ya que se comprueba que los fonos tienen una fuerte influencia de sus vecinos en el habla habitual de una persona.

- El porcentaje de reconocimiento es bastante parejo tanto en varones como en mujeres, lo cual indica que el sistema es bastante independiente de locutor, y que no hace mucha diferencia entre varones y mujeres.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Las diferencias entre frases indican que los fonos que las componen también influyen en el reconocimiento.
- Los porcentajes de reconocimiento más altos, en ambos sexos, pertenecen a aquellas personas que hablan más lentamente y que tienen un volumen de voz relativamente altom lo que le proporciona al módulo más material de análisism permitiéndole mayor precisión.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

# 6 RECOMENDACIONES



Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Construir bases de datos sobre diferentes contextos, tanto de voz como de texto [37,46]
- Cambiar de unidades, de monofonos (fonos aislados) a trífonos, los cuales en cuenta la influencia en el fono central de los fonos vecinos [24,39,40,41,42,43,47]
- Realizar un estudio sobre la pronunciación del español limeño, para poder construir un diccionario más preciso, e implementar árboles de decisión, los cuales limitan al módulo y reducen la carga computacional [38,48]
- Mejorar la gramática, dividiendo las palabras del vocabulario en categorías (sustantivo, verbo, artículo, etc), para limitar las opciones de las que dispone el sistema [32]
- Compilar un vocabulario con las palabras estrictamente necesarias, para reducir la probabilidad de que el sistema se confunda [32]
- Mejorar la robustez frente al ruido, empleando métodos tales como filtros digitales adaptivos [19,30]
- Utilizar un corpus de entrenamiento tomado en el entorno real en el que trabajará el módulo, con el mismo micrófono que se utilizará en la etapa de pruebas a ser posible.
- Desarrollar un segmentador automático, para acortar el tiempo necesario para preparar los corpus de voces, ya sea empleando como base un reconocedor de habla, modelos gaussianos o características temporales [20,21,22,26,49]
- Implementar todo o parte del módulo en hardware, para aumentar la velocidad de procesamiento [31]



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

- Implementar modelos híbridos, mezclando HMM con MLP (MultiLayer Perceptron, redes neuronales), los cuales combinan el modelado en el tiempo del primero con la clasificacion discriminativa de patrones del segundo [24,27,29,35,36,45] e investigar nuevas formas de utilizar los HMM, como los HMM usados para segmentar [50] o los modelos Gaussianos.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

# 7 BIBLIOGRAFIA



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[1] Huang, Xuedong, Acero, Alex y Hon, Hsiao-Wuen, “Spoken language processing : a guide to theory, algorithm, and system development”, Upper Saddle River, NJ, Prentice Hall PTR, 2001.

[2] O´Shaugnessy, Douglas, “Speech communications : human and machine”, 2da ed., New York, IEEE, 2000.

[3] Jurafsky, Daniel y Martin, James H, “Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition”, Upper Saddle River, NJ, Prentice Hall, 2000.

[4] Rabiner, Lawrence y Juang, Biing-Hwang, “Fundamentals of speech recognition”, Englewood Cliffs, NJ, Prentice Hall, 1993.

[5] Rabiner, Lawrence, “A tutorial on Hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, Vol.77, No 2, p. 257-286, Febrero 1989.

[6] San Román Denegri, Claudio, “Sistema verificador de locutor de texto variable basado en concadenación de modelos fonéticos”, Tesis de Licenciatura, Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería, Lima – Perú, 2003.

[7] Lam, Jorge y Mellado, Abel, “Diseño e implementación de un sistema de reconocimiento de locutor independiente del texto, usando modelos ocultos de MARKOW y cuantificación vectorial de ACW”, Tesis de Licenciatura, Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería, Lima – Perú, 2001.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[8] Young, Steve, Evermann, Gunnar, Hain, Thomas y otros, “The HTK Book (for HTK Version 3.2.1)”, Cambridge University Engineering Department, December 2002.

http://htk.eng.cam.ac.uk/docs/docs.shtml

[9] ELRA (European Language Resources Association), Catalogue of Language Resources.

http://www.elra.info/

[10] The SALA Project

http://www.sala2.org/

[11] NASA :: Intelligent Systems :: Clarissa

http://tc.arc.nasa.gov/projects/clarissa/index.php?ta=&gid=&pid=

[12] HOMEY demonstrator – Interactive spoken dialogue management system

http://www.openclinical.org/dm_homey.html

[13] Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory

http://groups.csail.mit.edu/sls/sls-blue-noflash.shtml

[14] OVIS – Openbaar Vervoer Informatie Systeem

http://lands.let.kun.nl/TSpublic/strik/ovis.html

[15] Huang, Chao, Xu, Peng, Zhang, Xin, Zhao, Shubin, Huang, Taiyi y Xu, Bo, “LODESTAR : A mandarin spoken dialogue system for travel information retrieval”, Eurospeech, 1999, p. 1159-1162.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[16] Turunen, M., Salonen, E-P., Hartikainen, M., Hakulinen, J., Black, W. J., Ramsay, A. Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gambäck, B., Cheadle, M., Olsson, F., Sahlgren, M. 2004. "AthosMail - a multilingual Adaptive Spoken Dialogue System for E-mail Domain", Proceedings of the 20th International Conference on Computational Linguistics, pp. 77-86, Ginebra, Suiza.

[17] Sistemas de diálogo basados en procesamiento del habla y multimodales.

http://www.ugr.es/~rlopezc/sistemas_dialogo.htm

[18] Corpus orales para la fonética y las tecnologías del habla en español

http://liceu.uab.es/~joaquim/language_resources/spoken_res/Corp_oral_esp.html

[19] Moreno, Pedro J. y Stern, Richard M. “Sources of degradation of speech recognition in the telephone network”, *Proceedings of the IEEE*, Vol. 1, pp. 109-112, Abril 1994, Adelaide, Australia.

[20] Demuynck, K. y Laureys, T., “A comparison of different approaches to automatic speech segmentation”, *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pp. 277-284, 2002

[21] Bonafonte, A., Nogueiras, A. y Rodríguez-Garrido, A., “Explicit segmentation of speech using Gaussian Models”, *ICSLP*, pp.1269-1272, 1996

[22] Wang, D., Lu, L. y Zhang, H. J., “Speech segmentation without speech recognition”, *Proceedings of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 468-471, Hong Kong, 2003.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[23] Ahadi, S.M., Sheikhzadeh, H., Brennan, R.L. y Freeman, G.H., "An efficient front-end for automatic speech recognition", *International Conference of Electronics, Circuits and Systems (ICECS2003)*, United Arab Emirates, 2003.

[24] Koreman, J., Barry, W. J., y Andreeva, B., "Relational phonetic features for consonant identification in a hybrid ASR system".

[25] Deng, Li, Acero, A., Plumpe, M., Huang, X., "Large-vocabulary speech recognition under adverse acoustic environments", *Proceedings of the International Conference on Spoken Language Processing*, Vol. 3, pp. 161-164, 2000

[26] Docío, L. y Mateo, C., "Segmentación automática de voz basada en modelos ocultos de Markov y características acústicas", *Procesamiento del Lenguaje Natural*, Revista nº 26, setiembre 2000

[27] Johansen, F., "A comparison of hybrid architectures using global discriminative training", *COST 249 meeting*, Madrid, 1995

[28] Schwartz, R. y Chow, Y. L., "The N-best algorithm : An efficient and exact procedure for finding the N most likely sentence hypotheses", *Proc. Int. Conf. Acoust. Speech, Sign. Proc (ICASSP)*, pp. 81-84, 1990

[29] Boite, J-M., y Ris, C., "Developement of a French speech recognizer using a hybrid HMM/MLP system", *European Symposium on Artificial Neural Networks*, ISBN 2-600049-9, pp. 441-446, 1999

[30] Entwistle, M., "The performance of automated speech recognition systems under adverse conditions of human exertion", *International Journal of Human-Computer Interaction*, pp. 127-140, 2003

[31] Melnikoff, S.J., Quigley, S.F. y Russell, M.J., "Implementing a simple continuous speech recognition system on an FPGA", *Proceedings of the 10th Annual IEEE*



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

*Symposium on Field-Programmable Custom Computing Machines (FCCM´02)*, 2002

[32] Young, S., "Large vocabulary continuous speech recognition: A review", *Proceedings of the IEE Workshop on Automatic Speech Recognition and Understanding*, pp. 3-28, 1995

[33] Everett, S., Wauchope, K. y Perzanowski, D., "Talking to a Natural Language Interface : Lessons Learned", http://www.aic.nrl.navy.mil/papers/1992/AIC-92-016.ps, Navy Center for Applied Research in Artificial Intelligence Naval Research Laboratory, 1992

[34] Ferreirosm J., Macas-Guarasa, J., Pardo, J. M., y Villarrubia, L., "Introducing multiple pronunciations in Spanish speech recognition systems", *Proceedings of ESCA Workshop ´Modeling Pronunciation Variation for Automatic Speech Recognition´*, pp. 29-34, 1998

[35] Cohen, M., Franco, H., Morgan, N., Rumelhart, D., Abrash, J., y Konig, Y., "Combining neural networks and hidden Markov models", *Proceedings of the DARPA Speech and Natural Language Workshop*, Harriman, NY, 1992

[36] Deroo, O., Ris, C., Malfrere, F. y Dutoit, T., "Hybrid HMM/ANN systems for speaker independent continuous speech recognition in French"

[37] Markhus, V., Gajic, B., Svarverud, J., Solbraa, L. E. y Johnsen, M., "Annotation and automatic recognition of spontaneously dictated medical records for Norwegian", *Proceedings of the 6th Nordic Signal Processing Symposium* – NORSIG 2004, Finlandia, 2004

[38] Yu, H. y Schultz, T., "Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition", EUROSPEECHm Ginebra 2003



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[39] Odell, J.J., “The use of context in large vocabulary speech recognition”, Ph.D. diss, University of Cambridge, 1995

[40] Beulen, K., Bransch, E. y Ney, H., “State tying for context dependent phoneme models”, *European Conf. On Speech Communication and Technology*, Rhodos, Greece, pp. 1179-1182, 1997

[41] Kershaw, D., “Phonetic context-dependency in a hybrid ANN/HMM speech recognition system”, Ph.D Thesis, Cambridge University Engineering Department, 1996

[42] Sixtus, A. y Ney, H., “Training of across-word phoneme models for large vocabulary continuous speech recognition”, 2002

[43] Shafran, Z. y Ostendorf, M., “Use of higher level linguistic structure in acoustic modeling for speech recognition”, *Proceedings of ICASSP*, Vol. 2, pp. 1021-1024, 2000

[44] Shafran, L., Ostendorf, M., y Wright, R., “Prosody and phonetic variability: Lessons learned from acoustic model clustering”, *Proceedings ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 127-131, 2001

[45] Neto, J., Martins, C. y Almeida, L., “A large vocabulary continuous speech recognition hybrid system for the Portuguese language”, *Proceedings of ICSLP´98*, Sydney, Australia, 1998

[46] Neto, J., Martins, C., Meinedo, H. y Almeida, L., “The design of a large vocabulary speech corpus for Portuguese”, *Proceedings of EUROSPEECH 97*, Rodas, Grecia, 1997

[47] Zhengm J., Franco, H. y Stolcke, A., “Modeling word-level Rate-of-Speech variation in large vocabulary conversational speech recognition”.



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

TESIS PUCP

[48] Sun, J. y Deng, L., “A phonological modeling system based on autosegmental and articulatory phonology”

[49] Essa, O., “Using prosody in automatic segmentation of speech”, *Proceedings of the ACM 36th Annual Southeast Conference*, pp. 44-49, 1998

[50] Ostendorf, M., Digalakis, V. y Kimball, O. A., “From HMMs to segment models : A unified view of stochastic models for speech recognition”, 1996



SOME RIGHTS RESERVED
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis