Universidad de las Ciencias Informáticas
Facultad #9.

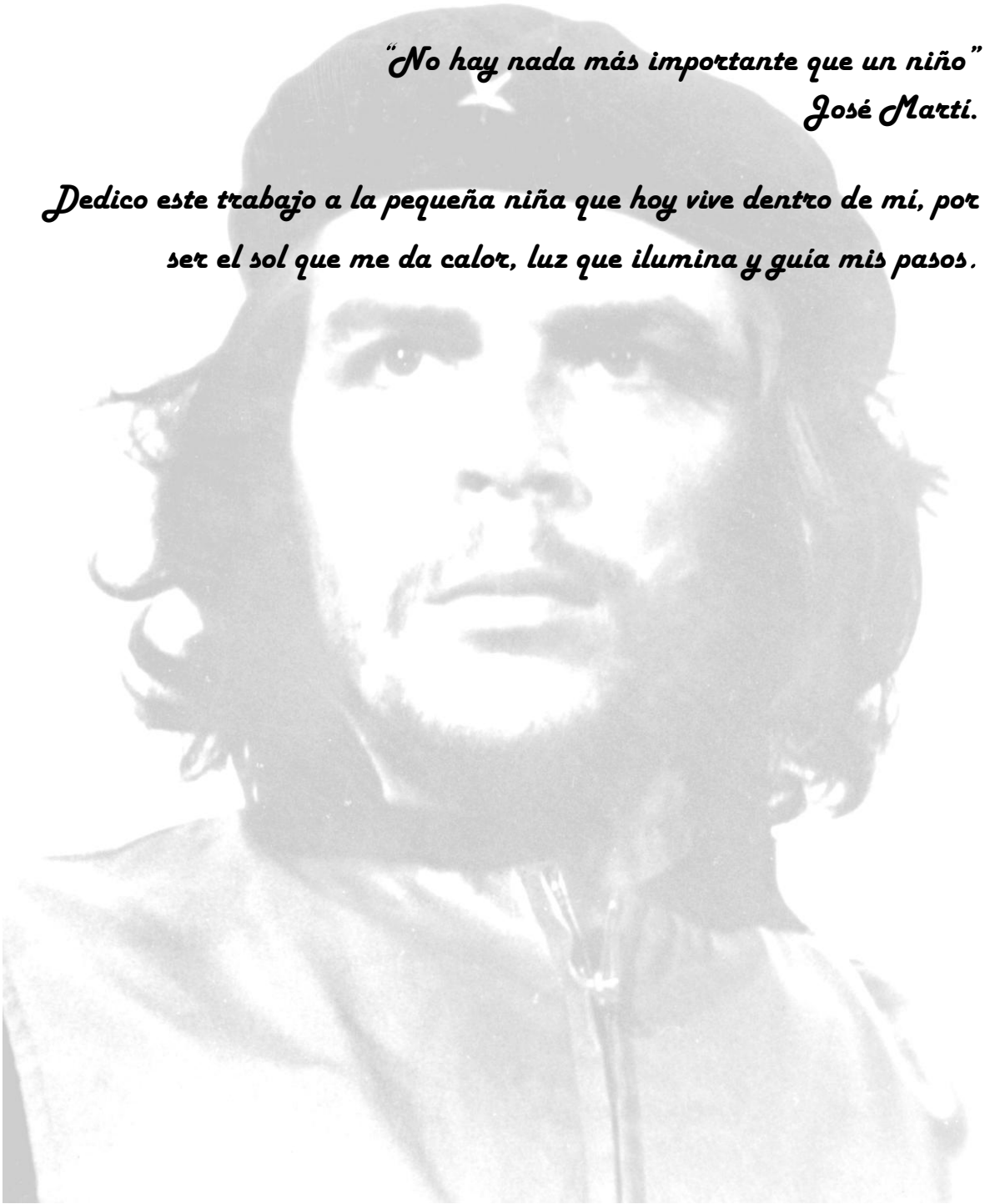

# Titulo: Conceptualización de un sistema informático para el reconocimiento y autenticación de personas por la voz.

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN INFORMÁTICA

Autor: Dianabel Sánchez Otero
Tutor: Ing. Aliosmi López Velazquez

Ciudad de la Habana 26/05/2010
Año 52 de la Revolución.

# DEDICATORIA



*"No hay nada más importante que un niño"*
*José Martí.*

*Dedico este trabajo a la pequeña niña que hoy vive dentro de mí, por ser el sol que me da calor, luz que ilumina y guía mis pasos.*

# AGRADECIMIENTOS

*La gratitud es el más legitimo pago al esfuerzo ajeno, además de necesaria es hermoso; por eso le agradezco:*

*A mis padres, por constituir el impulso necesario para hacer realidad mis sueños, por sus desvelos, dedicación y amor.*

*A nuestra Revolución Socialista que hizo posible que tengamos el derecho a formarnos como profesionales y realizarnos como protagonistas de su obra.*

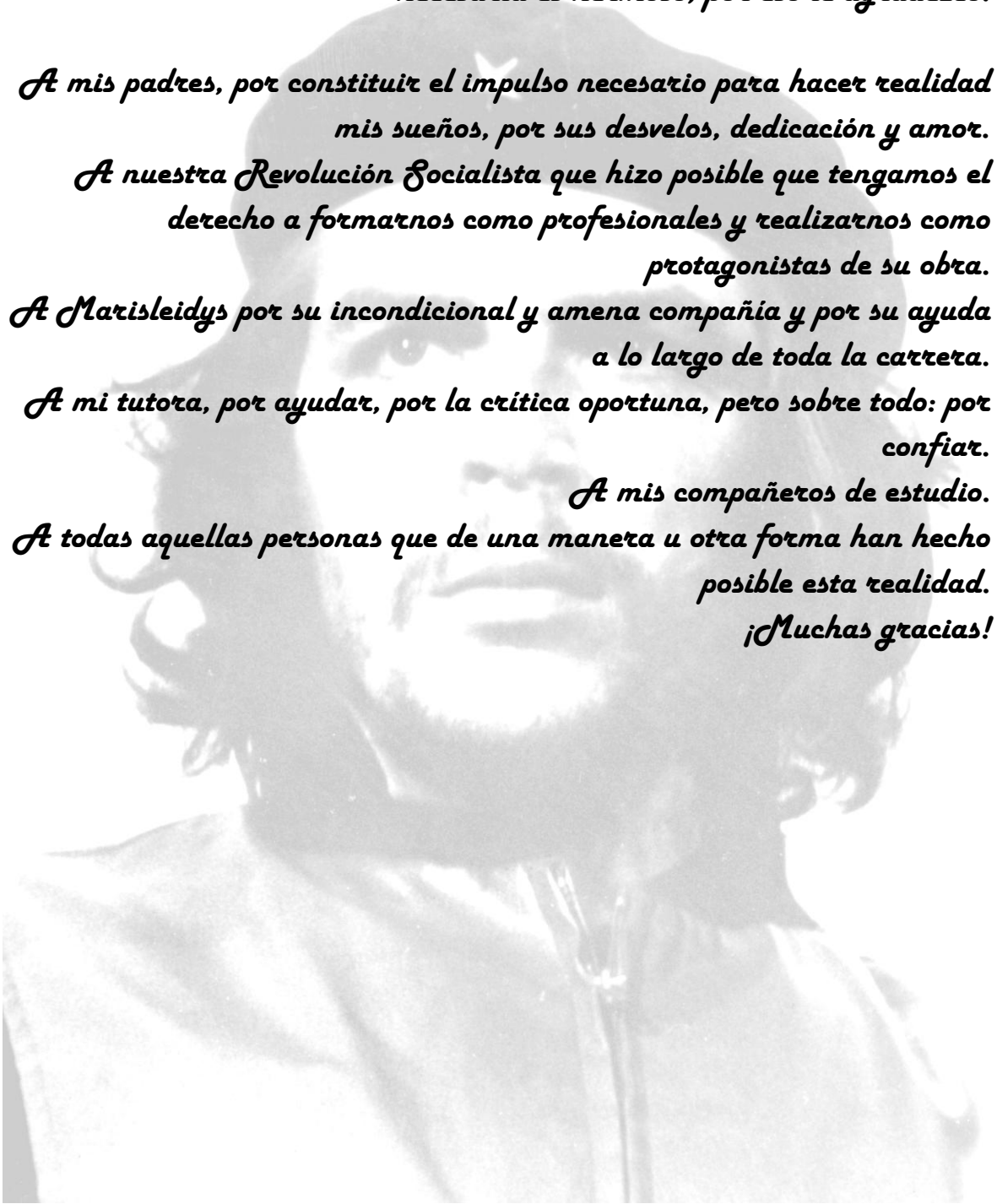
*A Marisleidys por su incondicional y amena compañía y por su ayuda a lo largo de toda la carrera.*

*A mi tutora, por ayudar, por la crítica oportuna, pero sobre todo: por confiar.*

*A mis compañeros de estudio.*

*A todas aquellas personas que de una manera u otra forma han hecho posible esta realidad.*

*¡Muchas gracias!*



# DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo al Facultad 9 de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los _26__ días del mes de ___5____ del año _2010___.

Dianabel Sánchez Otero
_____

Ing. Aliosmi López Velázquez
_____

# OPINIÓN DEL TUTOR

Título: Conceptualización de un sistema informático para el reconocimiento y autenticación de personas por la voz

Autor(es): Dianabel Sánchez Otero

El tutor del presente Trabajo de Diploma considera que durante el período que se evalúa la estudiante ha mostrado las cualidades que a continuación se detallan.

Muy alta independencia durante todo el desarrollo de su investigación, profundizando en los aspectos teóricos que permiten consolidar los conocimientos del tema en cuestión. Se ha esforzado para realizar la investigación a pesar de sus problemas personales y de salud; para lograr cumplir con las tareas planteadas para desarrollar la misma. En el documento se evidencia el resultado de su investigación así como el progreso que ha logrado desde sus inicios. El mismo está redactado con alto nivel científico y con la seriedad necesaria.

Por todo lo anteriormente expresado considero que la estudiante ha vencido las tareas planificadas, así como cumplido los objetivos propuestos; y propongo que se le otorgue al Trabajo de Diploma, la calificación de 5 puntos

Tutor: Ing. Aliosmi López Velázquez

_____ _____

Firma Fecha

# Datos de Tutor

Profesora de la asignatura de Ingeniería de Software I en el curso 07-08.

Analista de Sistemas en el Proyecto Productivo “Sistema Automatizado de Control y Gestión de Indicadores de Refinerías (SACGIR)”, Universidad de las Ciencias Informáticas (UCI), 2007. Asesora de Calidad en la producción de Software en el Polo Video y Sonido Digital de la Facultad #9, 2007 Administradora del Proyecto “Sistema de Monitoreo de Señales de Radio y Televisión” de la Facultad #9 desde 2007 hasta 2008.

Profesora de la asignatura Ingeniería de Software II en el curso 08-09. Profesora de la asignatura Ingeniería de Software I e Ingeniería de Software II en el curso 09-10.

**RESUMEN**

En estos últimos años y con el desarrollo de las TIC[1], ha surgido la necesidad de ampliar la seguridad en los sistemas informáticos que existen en todas partes del mundo. No solo estos sistemas necesitan de la seguridad sino también locales a los cuales pueda acceder un determinado grupo de personas que cumplan con un grupo de características que puedan identificarlas dentro de las demás.

Con este fin surgen los medios de autenticación biométricos, los cuales actualmente son muy difundidos por todo el mundo debido a la seguridad que los mismos representan para el reconocimiento de personas, utilizando algún rasgo físico o de comportamiento de estas.

Dentro de los sistemas biométricos que en la actualidad existen, se pueden destacar los de tratamiento de iris, los de huellas dactilares y los no menos importantes, SRAH[2], entre otros muchos.

De estos sistemas se estará tratando en el desarrollo del presente trabajo de diploma, haciendo énfasis principalmente en los SRAH, de los cuales se realizará una búsqueda para determinar las diferentes variantes que pueden ser aplicadas a la hora de realizar un sistema de este tipo y cuáles de ellas serían las más factibles de utilizar.

[1] **TIC** acrónimo de Tecnologías de la Informática y las Comunicaciones
[2] **SRAH** acrónimo de Sistemas de Reconocimiento Automático del Habla

## Tabla de ilustraciones



## Tabla de contenido

# INTRODUCCIÓN

Hoy en día se está viendo en todo el mundo el riesgo que corren las empresas debido al manejo inadecuado de la información confidencial; ocurriendo delitos como la malversación, el robo y otros efectos en estos datos que las hacen cada vez más vulnerables por lo que sus informáticos se ven en la imperiosa necesidad de concebir de manera más robusta la seguridad; restringiendo el acceso a los locales de las mismas. Para ello se basan en los sistemas biométricos, lo cual no es más que el análisis de uno o más rasgos físicos o de conducta propios de las personas para su reconocimiento.

En el mundo de la informática la autenticación biométrica consiste en la aplicación de tecnologías para medir, así como reconocer las características físicas conocidas como estáticas y las características de comportamiento conocidas como dinámicas en las personas, para el acceso a determinados locales. En el mundo existen múltiples ejemplos de características fisiológicas que sirven para este fin, como las huellas dactilares, las retinas, el iris, los patrones faciales, los patrones de venas de la mano o la geometría de la palma de la mano y dentro de las de comportamiento se incluyen la firma, el paso y el tecleo.

La voz se considera una mezcla de características físicas y de comportamiento, pero todos los rasgos biométricos comparten aspectos de ambas clasificaciones. Según estudios realizados se puede afirmar que ésta; cuenta con una facilidad de uso, gran precisión y aceptación, mientras que tiene un nivel de seguridad y estabilidad intermedio.

Es muy común que mucha gente piense que los sistemas de verificación de voz intentan reconocer lo que el usuario dice, pero lo que realmente se reconoce es una serie de sonidos y sus características para decidir si el usuario es quien dice ser. Para autenticar a un usuario utilizando un reconocedor de voz se debe disponer de ciertas condiciones para el correcto registro de los datos, como ausencia de ruidos, reverberaciones o ecos; idealmente, estas condiciones han de ser las mismas siempre que se necesite la autenticación.

El principal problema del reconocimiento de voz es la inmunidad frente a *replay attacks*, un modelo de ataques de simulación en los que un atacante reproduce (por ejemplo, por

medio de un magnetófono) las frases o palabras que el usuario legítimo pronuncia para acceder al sistema, una solución a este problema consiste en utilizar otro sistema de autenticación junto al reconocimiento de voz.

La voz representa un patrón muy usado actualmente en varias compañías internacionales como es el caso de la AGNITIO, la cual ha desarrollado una herramienta llamada BATVOX que ayuda a los peritos y policía científica a identificar con precisión las voces telefónicas grabadas comparándolas con otras grabaciones previas.

Cuba no se encuentra entre los más avanzados si de tecnología se trata, pero si ha venido realizando algunos proyectos que han contribuido en el ahorro económico para el país y aunque no se ha desarrollado aun ningún sistema de reconocimiento y autenticación de personas por voz, si se han estado realizando investigaciones sobre el tema y se espera que pronto se logren avances en este sentido.

En la Universidad de las Ciencias Informáticas se hace necesario un sistema de este tipo, ya que se cuenta con información confidencial en las locales de producción referente a la labor que se realiza en los mismos. Por lo que a dichas áreas debería tener acceso solamente el personal calificado; para evitar que la información sensible pueda ser malversada o usada con fines lucrativos.

Para ello el Centro de Señales Digitales y Geoinformática perteneciente a la Facultad 9 de dicha universidad, se ha dado a la tarea de investigar a profundidad sobre el tema, para posteriormente poder desarrollar el sistema de control de acceso y ponerlo en práctica. De ahí entonces surge el siguiente **problema a resolver** la inexistencia de la tecnología para gestionar los procesos de reconocimiento y autenticación automática de personas por voz en los locales de acceso restringido en la Universidad de las Ciencias Informáticas.

Partiendo de esta problemática se puede plantear que el **objeto de estudio** son los procesos de reconocimiento y autenticación automática por voz. Teniendo a su vez como **objetivo general** proponer un modelo arquitectónico para un Sistema Informático que permita manejar información sobre el proceso de reconocimiento y la autenticación por voz.

El **campo de acción** que abarca la investigación es la arquitectura de software para sistemas de reconocimiento y autenticación automática de personas por voz en la Universidad de las Ciencias Informáticas y se sostiene como **idea a defender** la realización de la investigación sobre los procesos de reconocimiento y autenticación por voz para definir la arquitectura que permitirá el desarrollo posterior de un Sistema Informático.

Con el fin de lograr la definición de un modelo arquitectónico que sirva de base para el desarrollo de un sistema informático de manera tal que solucione el problema planteado, se establecieron las siguientes tareas de la investigación:

- Caracterizar el estado del arte de los procesos de reconocimiento y autenticación automática por voz.
- Evaluar el contenido de la información obtenida acerca de estos procesos.
- Determinar posibles soluciones que permitan conceptualizar un posible sistema de autenticación por voz.
- Caracterizar los sistemas de identificación y autenticación por voz similares existentes dentro y fuera del país.
- Seleccionar las tendencias y tecnologías actuales más apropiadas para el desarrollo de la investigación.
- Proponer un modelo arquitectónico para el desarrollo de un sistema informático para la autenticación de personas por voz.

Para el desarrollo de la investigación se utilizaron métodos científicos entre los que se encuentran los teóricos y los empíricos; los cuales permiten identificar las características del objeto de estudio. Dentro de los teóricos se encuentra el histórico – lógico, con su utilización se logra realizar un estudio y evaluar el contenido de la información obtenida acerca del estado del arte de los procesos de reconocimiento y autenticación automática por voz en diferentes momentos históricos.

# CAPÍTULO 1: Fundamentación Teórica.

## Introducción

En el mundo actual los sistemas biométricos juegan un papel fundamental en el desarrollo de las tecnologías, porque tienen un gran desempeño dentro de la Seguridad Informática. En la Universidad de las Ciencias Informáticas se vienen desarrollando investigaciones acerca del tema, entre las que se pueden mencionar los sistemas de reconocimiento y autenticación por voz.

Los patrones de voz son uno de los rasgos biométricos más distinguidos de cada individuo, únicos de cada uno de ellos por lo que es aplicable dentro de la seguridad de una entidad para evitar el mal uso de la información. En este primer capítulo se estarán abordando los conceptos asociados al dominio del problema como son: voz, autenticación y biometría entre otros. De esta forma el lector puede interactuar mejor con el tema.

Posteriormente se comentará acerca del estado del arte de los sistemas de reconocimiento y autenticación por voz y se proporcionarán criterios relacionados con el tema.

## 1.1 Conceptos asociados al dominio del problema.

### 1.1.1. Autenticación

A medida que se ha desarrollado el mundo de la informática y perfeccionado las aplicaciones que actualmente se utilizan, se ha presentado la necesidad de gestionar la seguridad en las aplicaciones que son utilizadas diariamente. Para ello es preciso que los usuarios se autentiquen en las aplicaciones y de esta forma controlar el acceso a las mismas. **(Schapper, y otros, Diciembre 2004)**

La autenticación no es más que el acto de confirmación en el que algo o alguien se define como auténtico. La autenticación de un objeto puede significar la confirmación de

su procedencia, mientras que la autenticación de una persona a menudo consiste en verificar su identidad. La autenticación depende de uno o varios factores. **(Gish, 1999)**

La autenticación se hace necesaria debido al gran número de aplicaciones y usuarios que necesitan administrar sus propios contenidos en aquellos sistemas que son compartidos por más de un usuario a la vez, en los cuales cada cual solo pueden acceder única y exclusivamente a la información que sea desarrollada por el mismo.

### 1.1.2. Sonido

El sonido es un tipo de onda que se propaga únicamente en presencia de un medio que haga de soporte de la perturbación. Los conceptos generales sobre ondas sirven para describir el sonido, pero, inversamente los fenómenos sonoros permiten comprender mejor algunas de las características del comportamiento ondulatorio. **(Tribaldos, 1993)**

Se define sonido como la onda mecánica longitudinal que se propaga a través de un medio elástico producto de una fuente de vibración. Como todo movimiento ondulatorio, el mismo puede representarse como una suma de curvas sinusoides con un factor de amplitud, que se pueden caracterizar por las mismas magnitudes y unidades de medida que a cualquier onda, tales como: Longitud de onda (λ), frecuencia (f) o inversa del período (T) y amplitud. **(LIA, 2008)**

Teniendo en cuenta que un sonido cualquiera es la combinación de perturbaciones sonoras que difieren en los parámetros anteriormente expresados, siendo uno de los elementos indispensables para los procesos normales de la audición y el habla, **(Rayleigh, 1894)** se puede clasificar en:

- Audibles: Corresponde a las ondas sonoras cuyas frecuencias oscilan en un intervalo de 20 a 20 000 Hz.
- Infrasónicas: Todas aquellas ondas sonoras que tienen frecuencias por debajo del espectro audible por el oído humano.
- Ultrasónicas: Comprende las ondas sonoras que tienen frecuencias por encima del intervalo audible.

La expresión del carácter grave o agudo de un sonido está marcada por los valores de frecuencia, donde las perturbaciones de baja frecuencia se denominan sonidos graves y los de alta se conocen como sonidos agudos. El ruido es una combinación en mayor o menor medida de frecuencias con niveles diferentes en forma aleatoria, esta expresión también se conoce como sinónimo de contaminación acústica y no es más que un sonido agudo ya sea simple o complejo, con un marcado carácter inarmónico y una intensidad tan alta, que puede resultar incluso perjudicial para la salud humana. **(Mujica, 2008)**

Existen fuentes de ruido artificiales o generadores de ruido que emiten ruido blanco o rosa, son utilizados en acústica para realizar ciertas mediciones como aislamiento acústico, insonorización, reverberación, etc. Las ondas sonoras constituyen un flujo de energía a través de la materia, cuya intensidad determina una medida de la razón a la cual la energía se propaga a través de un cierto volumen espacial. En dependencia del tipo de señal mediante el cual son transmitidas puede ser clasificar en: Analógica o Digital. **(Miyara, 2000)**

EL sonido analógico no es más que un tipo de frecuencia representable por medio de una función matemática continua, en la que es variable su amplitud y periodo en función del tiempo. Por otra parte se define sonido digital como la codificación numérica de una señal eléctrica que representa una onda sonora, la cual consiste en una secuencia de valores enteros obtenidos mediante dos procesos principales: el muestreo y la cuantificación discreta de la señal eléctrica. **(LIA, 2008)**

El proceso de digitalización precisa la grabación de la altura actual de la onda de sonido a intervalos regulares. La longitud de estos intervalos se denomina tasa de muestreo y a dicho proceso se define como muestrear. A pesar de la pérdida inherente de información al convertir la información continua en discreta, inducida por los errores de cuantificación durante el proceso de digitalización, que impide que la señal digital sea exactamente equivalente a la analógica que la originó, es mucho más factible su uso, teniendo en cuenta que es más fácil de transmitir, almacenar o manipular, siendo menos sensible a las interferencias, además ante la pérdida de cierta cantidad de información, esta puede ser reconstruida gracias a los sistemas de regeneración de señales, detección y corrección de errores. **(Nauce Comunication,2010)**

Es importante resaltar que el oído humano es capaz de captar aproximadamente 44000 fonos por segundo, por tanto para que un sonido digital tenga la calidad requerida deberá estar basado en una frecuencia similar a los 44 KHz, aunque en la actualidad se han desarrollado tarjetas capturadoras de sonido profesionales que llegan hasta los 100 KHz, con el objeto de obtener un mayor número de puntos sobre la muestra, consiguiendo una calidad óptima, teniendo en cuenta que esta depende de la frecuencia del muestreo o número de mediciones que se hacen por segundo, incluyendo su resolución. **(Miyara, 1999)**

En la actualidad existen muchos formatos para los archivos de audio digital, los cuales se pueden dividir en dos categorías principales PCM (Modulación por Impulsos Codificados del inglés *Pulse-Code Modulation*) y comprimidos **(Feel the Noise, Octubre de 1999)**. Como bien aparece reflejado anteriormente el tamaño puede depender de la cantidad de canales que tenga el archivo y de la resolución en cuanto a la tasa de muestreo y profundidad.

- Los formatos PCM contienen toda la información que sale de un convertidor analógico a digital, sin ninguna omisión y por eso, tienen la mejor calidad. Dentro de esta categoría se encuentran los formatos WAV (*WAVEform audio format*) y AIFF (*Audio Interchange File Format* cuya traducción *es* Formato de Archivo de Intercambio de Audio). La diferencia principal que tienen estos formatos es el encabezado, alrededor de 1000 bytes al comienzo del archivo. **(Miyara, 2009)**

- Con el objetivo de minimizar el consumo de memoria física con respecto a los archivos PCM se concibieron los formatos de sonido comprimidos, como por ejemplo el MP3 (MPEG-1 Audio Layer 3), AAC (acrónimo de *Advanced Audio Coding*) y Ogg (formato de archivo contenedor multimedia, desarrollado por la Fundación Xiph.org). Ciertos algoritmos de compresión descartan información que no es perceptible por el oído humano para lograr que el mismo fragmento de audio pueda ocupar en memoria inclusive la décima parte o menos de lo que ocuparía de ser PCM. **(Simón, 2004)**

La reducción en tamaño implica una pérdida de información y por ello a los formatos de este tipo se les denomina formatos comprimidos con pérdida. Existen también formatos de archivo comprimido sin pérdida, dentro de los que se

cuentan el FLAC (*Free Lossless Audio Codec* o Códec libre de compresión de audio sin pérdida en idioma español)) y el *Apple Lossless Encoder*, cuyo tamaño suele ser de aproximadamente la mitad de su equivalente PCM. **(Mujica, 2008)**

### 1.1.3. Espectro de frecuencias

El análisis de frecuencia conduce a una representación gráfica diferente de la referencia matemática clásica de la amplitud en función del tiempo, utilizando esta vez la amplitud pero en función de la frecuencia. Dicha representación se suele denominar espectro o representación espectral. El análisis de espectros que se define como la transformación de una señal de la representación en el dominio del tiempo hacia el dominio de la frecuencia. **(Rochaix, y otros, 2010)**

Para describir de manera normalizada la repartición de las energías sonoras en el conjunto del espectro audible, este ha de ser recortado en bandas de amplitud y de apelación normalizadas, donde cada banda está designada por su frecuencia central **(Rayleigh, 1894)**. Las audiofrecuencias que conforman el espectro audible pueden ser subdivididas en función de los tonos:

- Tonos graves: Frecuencias bajas, correspondientes a las 4 primeras octavas, esto es, desde los 16 Hz a los 256 Hz.
- Tonos medios: Frecuencias medias, correspondientes a las octavas quinta, sexta y séptima, esto es, de 256 Hz a 2 kHz.
- Tonos agudos: Frecuencias altas, correspondientes a las tres últimas octavas, esto es, de 2 kHz hasta poco más 16 kHz.

La aplicación de técnicas de análisis espectral súper resolutivas mejora considerablemente el proceso de detección de los parámetros de cada sinusoide y permite focalizar el estudio evitando distorsionar toda la síntesis por una o dos componentes mal detectadas. **(Basso, 1999)**

Teniendo en cuenta que los ruidos que son sonidos complejos compuestos por la suma de varias emisiones de diferentes frecuencias, cuyas componentes son muy numerosas, es más factible su procesamiento basado en el análisis espectral, pues este último

permite determinar de forma precisa las frecuencias que componen un sonido cualquiera. **(Rayleigh, 1894)**

Uno de los factores más importantes es el proceso de estimación de las frecuencias y atenuación de las resonancias de la parte determinística del sonido, que configuran la distribución de polos de señal. Por este motivo, el uso de técnicas de estimación espectral súper resolutivas se centra sobre todo en aquellas basadas en la descomposición de valores singulares y auto valores de matrices de señal, con el objetivo de minimizar la influencia del ruido en el cálculo. **(Xuendong, Huang; Fileno, Alleva; Hon, Hsiao-Wuen; Mei-Y, 1992)**

### 1.1.4. Acústica

La acústica es una rama de la física interdisciplinaria que estudia el sonido, infrasonido y ultrasonido, es decir ondas mecánicas que se propagan a través de la materia (tanto sólida como líquida o gaseosa) (no se propagan en el vacío). A efectos prácticos, la acústica estudia la producción, transmisión, almacenamiento, percepción o reproducción del sonido. **(Rochaix, y otros, 2010)**

La acústica tiene sus orígenes en la antigua Grecia y roma. La misma se venía empleando desde hacía mucho tiempo antes con los comienzos de la música pero no es hasta el siglo VI A.C que Pitágoras decide estudiar cómo se componían los intervalos de la música. El mismo quería saber porqué algunos intervalos de música sonaban más bellos que otros. Después de esto Aristóteles (384-322 A.C) comprobó que el sonido consistía en contracciones y expansiones del aire. **(Davis, y otros, 1990)**

La comprensión de la física en los procesos acústicos avanzó rápidamente durante y después de la Revolución Científica. Galileo y Mersenne, descubrieron la forma de independizar todas las leyes de las cuerdas vibrantes, terminado así el trabajo que Pitágoras había comenzado 2000 años antes. Galileo escribió “Las ondas son producidas por las vibraciones de un cuerpo sonoro, que se difunden por el aire, llevando al tímpano del oído un estímulo que la mente interpreta como sonido”. **(Hernando Pericás, mayo 1993)**

### 1.1.5. Biometría

Se entiende por biometría a la tecnología de identificación basada en el reconocimiento de una característica física e intransferible de las personas, como por ejemplo, la huella digital. La biometría es un excelente sistema de identificación de la persona que se aplica en muchos procesos debido a dos razones fundamentales, la seguridad y la comodidad. **(TAPIADOR MATEOS, y otros, 2005)**

Entre las aplicaciones de identificación con biometría se puede encontrar:

- Control de acceso biométrico
- Control de presencia biométrico
- Lector biométrico
- Lector biométrico para integración

La biometría es un sistema automatizado que es capaz de funcionar parecido al funcionamiento del cerebro humano para reconocer a las personas según algún rasgo distintivo de las mismas. EL sistema para el reconocimiento de voz emplea la biometría física y de conducta con el objetivo de analizar patrones de habla e identificar al interlocutor. Para llevar a cabo esta tarea, el patrón creado previamente por el interlocutor, debe ser digitalizado y mantenido en una base de datos que generalmente es una cinta digital de audio. **(Espinosa Duró, 2004)**

En la actualidad los sistemas biométricos son unos de los más utilizados para llevar a cabo el control y reconocimiento de las personas. Aunque muchos han intentado atacar estos sistemas para así desacreditarlos, infiriendo que no constan de la seguridad que afirman sus creadores, sin embargo esto no se ha podido demostrar, la realidad es que son los más seguros actualmente debido a los medios que utilizan para la identificación. **(Universidad de La Rioja, 2001-2010)**

### 1.1.6. Audición y Habla

Tanto la audición como el habla implican procesos fisiológicos, permitiéndoles a los seres humanos estar al tanto de los eventos que no son perceptibles a sus otros órganos

sensoriales, así como señalar sucesos importantes en el ambiente, desempeñando una función crucial en la comunicación. **(Gonzalez, 1985)**

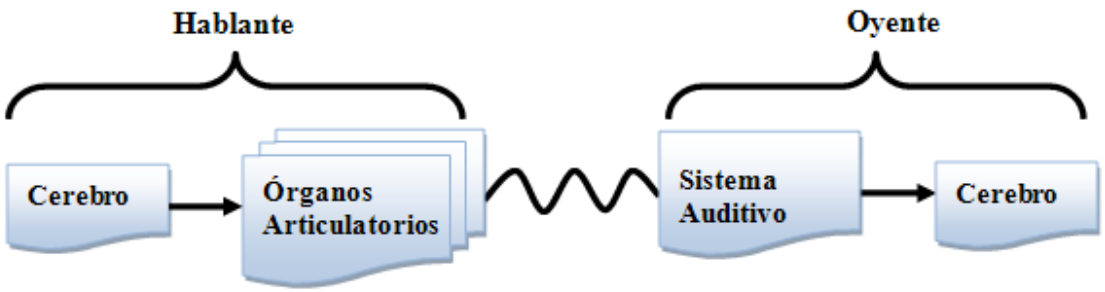



**Figura 1**: Sistema de comunicación humana

Teniendo en cuenta que la percepción sonora es el resultado de los procesos psicológicos que tienen lugar en el sistema auditivo, permitiéndole a los seres humanos interpretar los sonidos recibidos, implicando procedimientos fisiológicos derivados de la estimulación de los órganos de la audición **(Fletcher, 1995)**. Por tanto el sistema auditivo puede ser divido en dos partes fundamentales:

- Sistema auditivo periférico *(el oído)*, responsable de los procesos fisiológicos que captan el sonido y lo envía al cerebro.
- Sistema auditivo central *(nervios auditivos y cerebro)*, responsable de los movimientos psicológicos que conforman lo que se conoce como percepción sonora.

Es importante tener en cuenta que los sonidos del habla, al igual que todos los que se producen en la naturaleza no son tonos puros, sino complejas mezclas que se congregan en un espectro. Por tanto, el oído debe ser capaz no sólo de captarlos, sino de analizarlos y enviarlos al cerebro para que éste identifique los mensajes que portan. Donde juegan un papel fundamental los modos de audición, que no son más que modelos que permiten explicar cómo se produce el proceso de percepción sonora y de cómo se dota de significación a los sonidos. **(Gonzalez, 1985)**

Por otra parte no es común ver el procesamiento de la información auditiva separado en componentes individuales, es por ello que el mismo involucra simultáneamente el conjunto de habilidades que aparecen a continuación:

- Discriminación Sonora: Distinguir entre sonidos de diferente frecuencia, duración o intensidad.
- Localización: Ubicar la fuente sonora.
- Atención auditiva: Poner atención a las señales auditivas, especialmente al habla, durante un tiempo extenso.
- Figura Fondo Auditivo: Identificar a un hablante primario de un ruido de fondo.
- Discriminación Auditiva: Discriminar entre palabras y sonidos que son acústicamente similares.
- Cierre Auditivo: Comprender el mensaje completo cuando se pierde una parte.
- Armonización Auditiva: Sintetizar fonemas aislados que se encuentran "encapsulados" dentro de las palabras.
- Análisis Auditivo: Identificar fonemas o morfemas que se encuentran "encapsulados" dentro de las palabras.
- Asociación Auditiva: Identificar un sonido con su fuente.
- Memoria Auditiva, Memoria Secuencial: Almacenar y evocar estímulos auditivos de diferente longitud o número en el orden exacto.

El habla, como una manifestación sonora o acústica del lenguaje, se desarrolla a expensas de otros órganos y funciones anatómicas, como un sistema funcional sobre impuesto. Los sonidos producidos por las voz humana se reducen a unidades fonológicas formales de descripción que representan perturbaciones "ideales", abstractas y diferenciables de otras en una lengua determinada, los cuales se generan durante el proceso de expulsión del aire y se puede resumir en dos grandes apartados: la fonación y la articulación, aunque estos no deben ser vistos de forma independiente. **(Federico, 2007)**

Por tanto, se puede decir que la fonación no es más que la articulación de palabras, a través del proceso por el cual se modifica la corriente de aire procedente de los pulmones y la laringe en las cavidades supra glóticas (hace referencia a 4 cavidades: Faringe, Nasal, Oral y Labial) como consecuencia de los cambios de volumen y de forma de estas cavidades. El sistema fonador se vincula con otros sistemas y la interacción de estos es

una parte activa en la función fonadora que se regula por el sistema nervioso central y periférico. **(Syrdal, y otros, 1995)**

El control de las cuerdas vocales se produce mediante la participación de varios músculos y ligamentos situados en la laringe. Para ello basa su funcionamiento en el estiramiento o relajación de las mismas produciendo una mayor o menor frecuencia de la vibración, que es lo que se conoce como tonos altos o bajos del habla. **(Pelton, 1992)**

Todo lo expuesto anteriormente da origen al timbre de voz y la calidad vocal, sin embargo el sonido inducido es muy débil. Por ello debe ser amplificado, cuyo proceso tendrá lugar en los resonadores nasal, bucal y faríngeo, donde se producen modificaciones que consisten en el adelgazamiento para las altas frecuencia de ciertos fonos y el engrosamiento para las bajas. Por tanto el conjunto de las cavidades supra glóticas puede dividirse en tres partes: la faringe, la cavidad bucal y la cavidad nasal. **(Llorach, 1986)**

Modos de producción básicos de sonido de una lengua:

- Con la vibración de las cuerdas vocales se producen los sonidos tonales" o sonoros.
- Sin vibración de las cuerdas vocales y con interrupciones (totales o parciales) en el flujo de aire que sale de los pulmones que da lugar a los sonidos sordos.
- La combinación de vibración e interrupción, como las oclusivas sonoras.

### 1.1.7. Fonema

Se denomina fonemas al conjunto de abstracciones mentales o formales de los sonidos del habla, caracterizados por una especificación incompleta de rasgos fonéticos ya sean acústicos o articulatorios. En este sentido, un fonema puede ser representado por una familia o clase de equivalencia de ondas sonoras, técnicamente denominados fonos, que los hablantes asocian a un sonido específico durante la producción o la percepción del habla. **(Reccasens i Vives, 1993)**

Teniendo en cuenta lo planteado anteriormente los fonemas de una lengua no son sonidos, sino conjuntos de rasgos sonoros que los interlocutores se hayan adiestrado en producir y reconocer dentro de la corriente sonora del habla, los cuales se distinguen acústicamente por la envoltura del espectro, y particularmente por la frecuencia de los picos espectrales. Estos surgen de las resonancias del tracto vocal y se denominan formantes, identificados por medio de un identificador numérico, siendo el primer formante el de más baja frecuencia, donde el conjunto de formantes o rasgos sonoros conforman un espectro cuyo corpus o envoltura es en sí lo que constituye el fonema. **(Geomodeling Technology Corp, 2009)**

### 1.1.8. Formantes

Los formantes no son más que frecuencias que entran en resonancia en las cavidades nasales y orales, saliendo hacia el exterior como la información más importante del habla. Las principales diferencias entre las letras se basan principalmente en la ubicación de sus formantes. El cerebro humano analiza estas frecuencias y la relación que existe entre ellas, ante la duda decide que letra o vocal es la que se está pronunciando. Esto se puede definir como una operación que realiza nuestro cerebro para arribar a una conclusión final y entender lo que se ha escuchado. **(Boix, 1991)**

1. Técnicas de comprobación de patrones

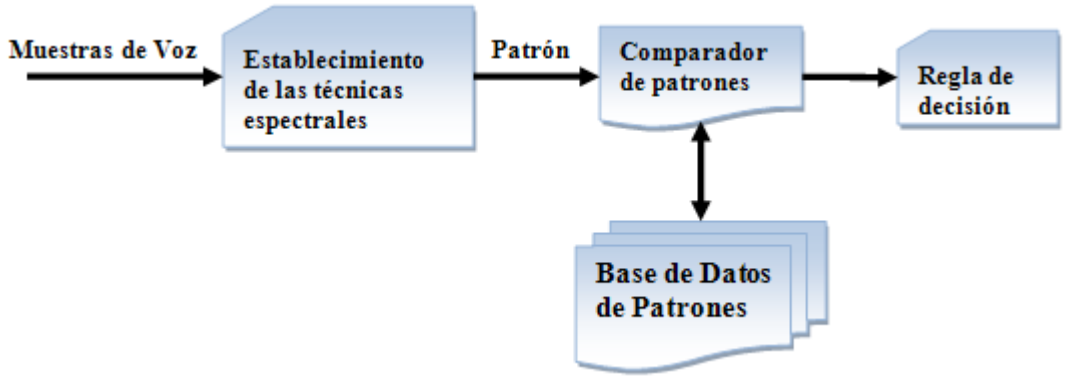



**Figura 2**: Reconocimiento del habla empleando técnicas de comprobación de patrones.

La principal ventaja que tiene esta técnica es que no es necesario descubrir características espectrales de la voz a niveles fonéticos, lo que evita tener que desarrollar etapas complejas de reconocimiento de formantes, de rasgos distintivos sonidos, tonos,

etc. Esta técnica es recomendada para trabajar con sistemas que estén compuestos de un conjunto no muy grande de palabras. **(Reccasens i Vives, 1993)**

2. Estudio basado en la posición de los formantes.

Para obtener una información detallada de los tres primeros formantes, la misma se puede ubicar en un espectrograma el cual nos da alguna información pero no tan detallada como se desea. Para ello se crean gráficas con los formantes que se tienen para identificar a la persona que habla, entonces se representa cada uno de los formantes en cada uno de los ejes coordenados formando gráficos en 3d que representan cada una de las personas que se identifiquen. **(Miyara, 2010 )**

### 1.1.9. Aparato Fonador

Se define como Aparato Fonador al andamiaje biológico del ser humano controlado por el sistema nervioso central en correlación con un conjunto de peculiaridades morfológicas y mecanismos fisiológicos, los cuales influyen de manera decisiva sobre las características del habla, tal y como se explicó en los acápites anteriores. Por eso ciertos rasgos fonéticos son comunes a todas las lenguas y muchos otros son altamente frecuentes. **(Boix, 1991)**

La emisión de un fonema exige la realización de determinadas maniobras neuromusculares, así como la generación de corriente de aire que debe ser modulada a diferentes niveles del aparato fonador. Las características neuromusculares, únicas en el hombre, hacen posible la emisión de sonidos que son utilizados como unidades informativas del lenguaje. **(Reccasens i Vives, 1993)**

Se considera importante tener en cuenta que la naturaleza de la estructura general del aparato fonador, que incluye los espacios geométricos del tracto vocal funciona como un sistema de resonancia, así como los movimientos de los articuladores permiten modular los sonidos fundamentales y sus armónicos. Teniendo en cuenta que la fisiología de la inervación y la estructura de la coordinación motora son las que permiten la realización de los movimientos necesarios para que el aparato fonador opere cambios permanentes en forma rápida pero precisa. **(Trubetzkoy, 1992 )**

### 1.1.10. Reconocimiento Automático de Habla

El Reconocimiento Automático de Habla o Reconocimiento Automático de Voz no es más que una rama de la Inteligencia Artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y sistemas de cómputos. El problema que se plantea en un sistema de RAH es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido. **(Martínez Celdrán, 2007 )**

El objetivo último del RAH es permitir la comunicación entre seres humanos y computadoras. Llevar a cabo un sistema de reconocimiento automático del habla consiste en la extracción de un conjunto de valores o parámetros que contengan la información más importante de la señal.

Por información relevante se entiende, aquella información acústica que pueda hacer que se reconozca el mensaje. Para un sistema de autenticación de habla constituye información relevante toda aquella que posibilita en determinada señal de voz reconocer quién es la persona que está intentando autenticarse en el sistema. **(Jesús Bernal Bermúdez)**

Para llevar a cabo el reconocimiento es necesario que el número de parámetros sea el menor posible. Esto viene dado por el tiempo de respuesta que pueda tener el sistema desde el momento en que se recibe la señal. Es demasiado costoso su procesamiento y por tal motivo se recomienda que la información sea la mínima. **(Jesús Bernal Bermúdez)**

Para lograr los objetivos de recursos consumidos y demás, la señal de voz es limitada en banda y se digitaliza. Como valores más frecuentes para la digitalización de las señales recibidas se utilizan frecuencias que se encuentren entre 3,7 KHZ y 5 KHZ aunque por lo general se muestra en una frecuencia entre 8 y 10 KHZ. A continuación se divide la señal

en segmentos de duración fija un valor típico es 30ms[3] solapados entre sí. Para cada mensaje se hace una comprensión de datos que consiste en un análisis de las frecuencias de la señal dando como resultado un conjunto respectivo de parámetros. **(Frega, y otros, 2000)**

Como se decía anteriormente los sistemas RAH generalmente reciben una determinada información. Esta información es vista generalmente como un problema de codificación en el cual se asume que la señal habla lleva implícito un mensaje codificado que consiste en una secuencia de símbolos. Dichos símbolos pueden representar fonos, sílabas, palabras o cualquier tipo de unidad sonora. **(Llorach, 1986)**

Por tal motivo es necesario dada la señal acústica los símbolos que conforman la misma. Para disminuir la enorme variabilidad de la señal acústica y permitir el tratamiento estadístico de la misma esta es parametrizada convirtiéndola en una secuencia de vectores equiespaciados llamados vectores acústicos. Para ello se calcula entonces la cantidad de símbolos codificados en la señal de la siguiente manera. **(Adriana Becerra, August 3, 2009)**

$$W = argmax P(W / Y)$$

Donde Y = y1.....yt es la secuencia de vectores acústicos también llamada secuencia de observación y W es la secuencia de símbolos correspondientes a dicha secuencia de observación. En los sistemas de reconocimiento automático de habla se tiene como secuencia de símbolos frases en las cuales se buscan dentro de ellas los símbolos que más probabilidades tengan de observar la secuencia de vectores acústicos y para esto tendremos entonces dos problemas a resolver. **(Adriana Becerra, August 3, 2009)**

La determinación del P (W/Y) o determinación del modelo estático. Esto se realiza a partir de grandes bases de datos de emisión de habla las cuales son convertidas con anterioridad en vectores acústicos y usadas para el entrenamiento del modelo estático. Entre los modelos estáticos el más utilizado en los últimos años ha sido el de los Modelos Ocultos de Markov (HMM) **(Bourlard, May 1995)**, el cual con diversas variantes ha

[3] **MS: Mili Segundos**

demostrado ser el mejor de todos los modelos existentes en lo que al procesamiento de habla respecta.

Hallar la fase óptima W que maximiza la probabilidad del modelo. Este procedimiento se realiza mediante la técnica de búsqueda de Viterbie **(Bourlard, May 1995)** que permite encontrar en forma más eficiente la frase más probable de acuerdo a la secuencia de vectores acústicos observados. **(Adriana Becerra, August 3, 2009)**

## 1.2. Objeto de estudio

### 1.2.1. Descripción General

La seguridad Informática es indispensable en cualquier empresa o entidad por muy grande o pequeña que sea, ya que ésta siempre va a contar con información confidencial que debe ser protegida, pero ¿Qué es la Seguridad Informática? La Seguridad Informática es un conjunto de técnicas encaminadas a obtener altos niveles de seguridad en los sistemas informáticos. Además, la seguridad informática precisa de un nivel organizativo por lo que se puede decir que:

Sistema de Seguridad = TECNOLOGÍA + ORGANIZACIÓN

Si bien es cierto que todos los componentes de un sistema informático están expuestos a un ataque (hardware, software y datos) son los datos y la información los sujetos principales de protección de las técnicas de seguridad. La seguridad informática se dedica principalmente a proteger la confidencialidad, la integridad y disponibilidad de la información.

A lo largo de estos últimos años los sistemas Biométricos se han vuelto indispensables en la Seguridad Informática de una entidad, ya que cada vez son más los números pin [4]y claves de acceso que se pueden confundir u olvidar fácilmente y más las tarjetas magnéticas que pueden ser robadas o se pueden extraviar . Por esto es que son más fáciles y a la vez más fiables los sistemas biométricos ya que estos son rasgos físicos o del comportamiento propios de cada individuo que están con ellos a cada momento y en

[4] **PIN: (***Personal Identification Number***)** Número de Identificación Personal

cada lugar en que se encuentren y no están a expensas de ser olvidados, extraviados o a que se los roben.

Los dispositivos biométricos tienen tres partes principales; por un lado, disponen de un mecanismo automático que lee y captura una imagen digital o analógica de la característica a analizar. Además disponen de una entidad para manejar aspectos como la compresión, almacenamiento o comparación de los datos capturados con los guardados en una base de datos (que son considerados válidos), y también ofrecen una interfaz para las aplicaciones que los utilizan.

El proceso general de autenticación sigue unos pasos comunes a todos los modelos de autenticación biométrica: captura o lectura de los datos que el usuario a validar presenta, extracción de ciertas características de la muestra (por ejemplo, las minucias de una huella dactilar), comparación de tales características con las guardadas en una base de datos, y decisión de si el usuario es válido o no.

Es en esta decisión donde principalmente entran en juego las dos características básicas de la fiabilidad de todo sistema biométrico (en general, de todo sistema de autenticación): las tasas de falso rechazo y de falsa aceptación. Por tasa de falso rechazo (*False Rejection Rate*, FRR) se entiende la probabilidad de que el sistema de autenticación rechace a un usuario legítimo porque no es capaz de identificarlo correctamente.

Por tasa de falsa aceptación (*False Acceptance Rate*, FAR) la probabilidad de que el sistema autentique correctamente a un usuario ilegítimo; evidentemente, una FRR alta provoca descontento entre los usuarios del sistema, pero una FAR elevada genera un grave problema de seguridad: se está proporcionando acceso a un recurso a personal no autorizado a acceder a él.

El reconocimiento y autenticación por voz no es más que un sistema que tiene guardado en una base de datos características y rasgos de la voz de una cierta cantidad de individuos, con el objetivo de obtener un patrón comparativo en el instante de la autenticación momento con la que se encuentra guardada en la base de datos y así

poder verificar si la persona es realmente quien dice ser y si cuenta con los permisos o no.

Hoy en día se han venido desarrollando en todo el mundo sistemas de reconocimiento y autenticación por voz y aunque estos son bastante fiables, como se decía en la introducción presentan algunas irregularidades como por ejemplo:

- El principal problema del reconocimiento de voz es la inmunidad frente a *replay attacks*, un modelo de ataques de simulación en los que un atacante reproduce (por ejemplo, por medio de un magnetófono) las frases o palabras que el usuario legítimo pronuncia para acceder al sistema. Este problema es especialmente grave en los sistemas que se basan en textos preestablecidos.

- Otro grave problema de los sistemas basados en reconocimiento de voz es el tiempo que el usuario emplea hablando delante del analizador, al que se añade el que éste necesita para extraer la información y contrastarla con la de su base de datos; aunque actualmente en la mayoría de los sistemas basta con una sola frase, es habitual que el usuario se vea obligado a repetirla porque el sistema le deniega el acceso.

- Una simple congestión hace variar el tono de voz, aunque sea levemente, y el sistema no es capaz de decidir si el acceso ha de ser autorizado o no; incluso el estado anímico de una persona varía su timbre. A su favor, el reconocimiento de voz posee la cualidad de una excelente acogida entre los usuarios, siempre y cuando su funcionamiento sea correcto y éstos no se vean obligados a repetir lo mismo varias veces, o se les niegue un acceso porque no se les reconoce correctamente.

### 1.2.2. Situación Problemática

En la Universidad de las Ciencias Informáticas existen muchas áreas de producción donde los estudiantes de conjunto con profesores y especialistas realizan un sin número de proyectos sumamente importantes tanto para el país como para entidades internacionales. Dichas áreas cuentan con información sumamente confidencial, las cuales están a expensas de ser sustraídas, cambiadas o usadas con fines lucrativos. Un

ejemplo de estos proyectos son el ERP, Primicia, Plataforma de Trasmisión Abierta para Radio y Televisión, entre otros muchos que no han sido mencionados.

Por eso es que se le ha dado la tarea al Centro de Señales Digitales perteneciente a la Facultad 9 de investigar sobre las posibles soluciones para implantar una seguridad más robusta en estas áreas de producción, con el fin de evitar todos estos malos usos de la información, es decir que solamente tengan acceso a ellas el personal calificado.

Una posible solución a este problema sería una investigación exhaustiva acerca de los procesos de reconocimiento y autenticación automática de personas por voz, para su posterior implementación, con el fin de implantar este sistema en las áreas de acceso restringido de la universidad.

Estos sistemas son seleccionados teniendo en cuenta que la voz es uno de los rasgos biométricos que en el proceso de implementación de un sistema para su reconocimiento no llevaría mucho tiempo y el hardware no sería muy costoso, además que cuenta con una facilidad de uso, una precisión y una aceptación bastante alta, mientras que tiene un nivel de seguridad y una estabilidad intermedia.

Esto se debe principalmente a que los sistemas de reconocimiento automáticos de habla utilizan solamente un micrófono en la parte del *hardware,* lo cual no es tan costoso.

## 1.3. Estado del arte de los procesos de reconocimiento y autenticación automática por voz.

Mucho antes del desarrollo del procesamiento de las señales modernas científicos de todo el mundo vertían todos sus esfuerzos por crear máquinas que produjesen habla humana. En 1779, el investigador danés Christian Gottlieb Kratzenstein, que trabajaba en esa época en la Academia Rusa de las Ciencias, construyó modelos del tracto vocal que podían producir las cinco vocales largas. Wolfgang von Kempelen de Vienna, Austria, describió en su obra *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine* ("mecanismo del habla humana con descripción de su máquina parlante", J.B. Degen, Wien) una máquina accionada con un fuelle. **(Federico, 2007)**

Esta máquina tenía, además, modelos de la lengua y los labios, para producir consonantes, así como vocales. En 1837 Charles Wheatstone produjo una 'máquina parlante' basada en el diseño de von Kempelen, y en 1857 M. Faber construyó la máquina 'Euphonia'. El diseño de Wheatstone fue resucitado en 1923 por Paget.

En los años 30, los laboratorios Bell Labs desarrollaron el VOCODER, un analizador y sintetizador del habla operado por teclado que era claramente inteligible. Homer Dudley refinó este dispositivo y creó VODER, que exhibió en la Exposición Universal de Nueva York de 1939.

Los orígenes del RAH se remontan a los años 40, momento en el que se desarrollan los primeros espectrógrafos, que permitían observar el espectrograma de una señal, dándole seguimiento a la evolución temporal de la energía en las distintas bandas de frecuencia, dato que podía servir para caracterizar y reconocer la voz humana. Como consecuencia la mayoría de los trabajos de la época se basaban en dispositivos analógicos que obtenían información acerca del contenido espectral de las señales, y utilizaban como criterio de clasificación las frecuencias de resonancia de las vocales. **(Federico, 2007)**

El primer dispositivo automático de reconocimiento, fue desarrollado en 1952 en los laboratorios Bell por Davis, Bidulph y Balashek, los cuales idearon un sistema totalmente electrónico, capaz de discriminar con cierta precisión los dígitos ingleses pronunciados de forma aislada por un mismo locutor, basándose en identificar las frecuencias de resonancia de la parte vocálica de los dígitos.

De una forma u otra en la década del 50 se comenzaron a desarrollar un grupo de investigaciones que constituirían las bases tecnológicas para lo que se conoce hoy en día por RAH, tal es el caso del proyecto de reconocimiento fonético llevado a cabo en la University College in England, el reconocedor de vocales independiente del hablante desarrollado en los laboratorios MIT[5] Lincoln, así como el reconocimiento de 10 sílabas mono-locutor, cerca de 1956 en los laboratorios RCA, la cual consistía en el reconocimiento de dichas sílabas mediante distancias espectrales obtenidas a partir de un banco de filtros analógico. **(Federico, 2007)**

[5] **MIT** acrónimo de Instituto de Tecnología de Massachusetts

Es importante destacar que el primer sistema de síntesis computarizado fue creado a finales de esta década y el primer sistema completo texto a voz se finalizó en 1968. Desde entonces se han producido muchos avances en las tecnologías usadas para sintetizar voz.

En los años 60 la comunidad científica internacional centró sus esfuerzos en la publicación de ideas o modelos conceptuales para el reconocimiento de patrones, permitiendo dar paso a los primeros trabajos que emplearon medios informáticos aplicados al RAH, provocando de esta forma una explosión de proyectos principalmente basados en el reconocimiento de palabras aisladas, con la impresión optimista de poder extrapolar los resultados y llegar en poco tiempo a sistemas capaces de reconocer cualquier frase pronunciada por cualquier locutor de manera continua.

Estos utilizaban técnicas de programación dinámica para comparar la secuencia de vectores de entrada, mediante alineamiento temporal no lineal DTW[6], con los patrones de las palabras del diccionario. Además de RCA (*Radio Corporation of America*) y AT&T (*American Telephone and Telegraph*), entran en escena los laboratorios japoneses de NEC (*Nippon Electric Company*), a los que se suman los trabajos realizados en la CMU[7], que continuarán hasta nuestros días. **(Llorach, 1986)**

En la década de los 70 se inició una proliferación de artículos científicos sobre sistemas de reconocimiento por voz y al mismo tiempo que se investigaba en temas relativos a reconocimiento del habla y síntesis de voz, pues ya había sido parcialmente solucionada la identificación de palabras aisladas, permitiendo de esta forma centrar todos los esfuerzos en el discurso continuo, donde se perfilaban dos aproximaciones dicho problema: los MEE[8] y los SBC[9]

En esta época se iniciaron una serie de ambiciosos proyectos muy prometedores, tal es el caso de ARPA-SUR[10] iniciado en 1971 y financiado por el Departamento de Defensa de los Estados Unidos de América. Por otra parte el consorcio de IBM decide no mantenerse al margen del desarrollo tecnológico en esta rama, apostando por los

---

6 **DTW** acrónimo de Dynamic Time Warping

7 **CMU** acrónimo de Carnegie Mellen University

8 **MEE** acrónimo de Modelos Estructurales Estocásticos

9 **SBC** acrónimo de Sistemas Basados en el Conocimiento

10 **ARPA-SUR** acrónimo de Advanced Research Projects Agency - Speech Un derstanding System

sistemas estadísticos-probabilísticos basados en el aprendizaje inductivo, conformando para ello un grupo de reconocimiento del habla, que ataca principalmente varias direcciones sobre grandes vocabularios.

Con el objetivo de de obtener sistemas independientes del locutor en AT&T se continúan las investigaciones con palabras aisladas y DTW, para lo cual se desarrollan algoritmos de agrupamiento de muestras que generen patrones robustos.

Es importante destacar que aunque los ambiciosos objetivos emprendidos por estos proyectos no llegaron a alcanzarse, sus aportes si contribuyeron de forma muy notable a un mejor conocimiento de los mecanismos del habla, los problemas y las limitaciones relacionados con el reconocimiento automático del mismo, derivados de la complejidad de dichos mecanismos, así como la toma de conciencia sobre la verdadera magnitud del problema planteado y la necesidad de una mayor investigación en el tema. **(Federico, 2007)**

La década de los 80 se caracterizó principalmente por la expansión y refinamiento de los algoritmos para el habla continua y grandes vocabularios, así como una explosión de los métodos estadísticos. Demostrada en los primeros años de esta década la ineficacia de los SBC (*Security Bank Corporation*), se invierte todo el esfuerzo en desarrollar sistemas capaces de extraer conocimiento de forma inductiva, tomando como base un conjunto de muestras.

A partir de investigaciones realizadas por IBM (*International Business Machines*), se comenzó a implementar la modelización acústica basada en HMM[11], centrada en elementos discretos y continuos, permitiendo optimizar los algoritmos de aprendizaje para entrenar los sistemas a partir de grandes bases de datos. Por tanto se mejoran también los sistemas de DTW para el reconocimiento de palabras conectadas, más concretamente se desarrollan algoritmos de búsqueda eficientes con los que determinar la sucesión óptima de patrones para una secuencia de vectores acústicos. **(Federico, 2007)**

Mediada la década de los 80 se presenta la aproximación conexionista como alternativa a la aproximación estadístico-probabilística y es ahí donde intervienen por primera vez las

[11] **HMM** acrónimo de Hidden Markov Models

redes neuronales artificiales, más conocidas como ANN[12], las cuales comparten con los HMM su carácter inductivo, que no es más que definir el aprendizaje a partir de muestras, pero sus configuraciones clásicas como los perceptrones multicapa.

Los perceptrones multicapa no son capaces de representar fenómenos dinámicos como la señal de voz, por lo cual tuvieron que desarrollarse arquitecturas recursivas específicas, con el objetivo de superar estas limitaciones. Otros autores han optado por configuraciones híbridas en las que un MLP[13] es utilizado para estimar las probabilidades de emisión de un HMM. **(Llorach, 1986)**

En los años 90 el vertiginoso desarrollo tanto del hardware como del software, permitió elevar los RAH un escalón más alto de lo que se había concebido hasta la fecha, tal es el caso de los sistemas de dictado automático y la integración entre el reconocimiento de voz y el procesamiento del lenguaje natural. Esto fue lo que sentó las bases para que 10 años después diera lugar a la integración de aplicaciones por telefonía móvil y sitios de Internet dedicados a la gestión de reconocimiento de voz, más conocido como Voice Web Browsers, así como la aparición del estándar VoiceXML. **(Llorach, 1986)**

Actualmente la investigación se concentra, en dos vertientes principales. La primera consiste en el mejoramiento del rendimiento de la modelización acústica a partir de la generación automática de unidades acústicas contextuares y entrenamiento discriminativo de los modelos. La segunda vertiente centra los esfuerzos en la integración de niveles de conocimiento superiores o estrategias de búsqueda heurísticas en grandes autómatas que representan modelos del lenguaje.

También se está invistiendo un gran esfuerzo en diseñar y adquirir grandes bases de datos para el entrenamiento de sistemas de reconocimiento de discurso continuo. **(Federico, 2007)**

## 1.4. Análisis de soluciones existentes

[12] **ANN** acrónimo de Artificial Neural Networks
[13] **MLP** acrónimo de MultiLayer Perceptron

Algunas de las cualidades más importantes de la voz son su naturalidad y versatilidad, que hacen de ella un método biométrico agradable para el usuario y adecuado para un buen número de aplicaciones. Es por ello que en la actualidad ha habido una explosión en el desarrollo de aplicaciones basadas en estos métodos.

Las empresas que lideran el mercado en el campo de las tecnologías para portales y sistemas de autenticación de voz, son Nuance, SpeechWorks, y Comverse. SpeechWorks, esta última consta de la tecnología necesaria para el reconocimiento del habla, tales como SpeechWorks 6.5, un TTS[14] denominado Speechify, y verificación de huella local como SpeechSecure, por tan solo citar unos ejemplos.

Nuance provee un reconocedor de voz llamado Nuance 7.0, uno de huella vocal nombrado Verifier 2.0, y se encuentra actualmente en función de preparar un TTS. Por otro lado, Commverse Network Systems comercializa una plataforma de operadores de portal de voz personal, conocida como Tel@GO 2.0, y un sistema de mensajería de voz sobre SMS, llamado VoiSMS. **(Rabiner, 1991)**

Otra de las aplicaciones que deben de tenerse en cuenta durante esta investigación es precisamente BioCloser, el cual no es más que un sistema de control de acceso biométrico. Admite distintos niveles de seguridad en función de la aplicación y utiliza la voz como método de autenticación. Este requiere generalmente de un proceso de entrenamiento previo en el que se creará un modelo o patrón de voz y un umbral de reconocimiento.


En el momento del reconocimiento, se compara una secuencia de voz con un modelo determinado y se comprueba si el umbral es superado. En ese caso, se considera que el usuario es quien dice ser. Los patrones de voz pueden ser mejorados con nuevos datos procedentes de los usuarios.

El sistema permite realizar una gran variedad de controles, sobre los accesos que tienen lugar en el sistema. El administrador es informado de todas las entradas por fecha, usuario, hora, etc. Además, es posible establecer restricciones de uso a usuarios o

[14] **TTS** acrónimo de Text to Speech

grupos de usuarios, atendiendo a criterios horarios. Esta herramienta fue desarrolladla por SeMarket es una empresa española fundada en 1999 especializada en el desarrollo de productos y servicios de seguridad en las áreas de identificación y verificación de identidades, control de acceso, y firma electrónica. Además forma parte de ABIE[15], SEAF[16], Plataforma eSec, Open Grid Forum y Terena Association**. (Puig, 1997)**

Sin embargo todos estos software, necesitan correr sobre una plataforma, y es ahí donde entran los servidores de voz. Eso es lo que suministra Intel Dialogic, cuya plataforma hardware es denominada Voice Portal Platform. Se compone de un soporte de Intel más una serie de tarjetas Intel Dialogic, que proporcionan el acceso de línea a la red telefónica y los recursos sobre los que se van a apoyar los suministradores de tecnología de reconocimiento de lenguaje natural.

Se trata de un sistema de base escalable, lo suficientemente flexible como para poder desarrollar nuevos servicios sobre la misma. Su capacidad de crecimiento para soportar el incremento del tráfico es muy buena. Dadas sus características, Óbice Portal Platform lo mismo sirve para un portal de voz, una aplicación de banca telefónica o un servicio de horóscopos. **(Puig, 1997)**

Los sistemas que fueron analizados en su mayoría son sistemas que se utilizan en ambiente web para el reconocimiento de personas utilizando llamadas telefónicas. En nuestro caso el sistema que se necesita no necesita ser web aparte de que como los algoritmos que se utilizan son demasiado pesados se aumentaría considerablemente el uso de memoria lo que traería consigo una gran pérdida de eficiencia en el sistema que se desea desarrollar.

[15] **ABIE** acrónimo de Asociación de Biometría Informática Española
[16] **SEAF** acrónimo de Sociedad Española de Acústica Forense

## Conclusiones parciales

En el capítulo que recién termina fueron registrados un conjunto de conceptos asociados al dominio del problema los cuales están muy relacionados con el tema que se lleva a cabo como son biometría, entre otros. Estos conceptos reflejan una mejor visión de cómo es que funcionan los sistemas biométricos que existen en la actualidad.

Además se precedió a la descripción exhaustiva del objeto de estudio y con el mismo se analizaron diferentes herramientas que son utilizadas en el mundo que llevan a cabo el control de acceso biométrico utilizando diferentes métodos biométricos entre los que resalta el uso de la voz que es el tema en cuestión.

# CAPÍTULO 2: Tendencias y Tecnologías

## Introducción

En este capítulo se abordarán temas relacionados con las diferentes tecnologías que en el mundo de hoy existen con respecto a los sistemas de Reconocimiento Automático de Habla (RAH). Además de esto se verán los modelos ocultos de Markov para los procesos de búsqueda de información en el momento de autenticar a los usuarios.

## 2.1. Modelos.

Para desarrollar un sistema RAH actualmente son utilizados frecuentemente los Modelos Ocultos de Markov (HMM) por sus siglas en ingles, estos constituyen la técnica más utilizada en todos los laboratorios del mundo en estos temas.

Un HMM resulta de la composición de dos procesos estocásticos, en las que la secuencia de unidades de reconocimiento subyacente y las observaciones acústicas están modeladas como procesos de Markov. Dicho de otro modo, se modela, por una parte, la secuencia temporal y, por otra, los eventos acústicos que se producen en dicha secuencia. Debido a que no es de sumo interés no se describirán aquí las fórmulas que se utilizan en los HMM sino que se remitirán a los lectores a**.** Para una descripción más personalizada en el funcionamiento de estos modelos se puede referir a **(Sanchis)**

La evidencia resultante de su uso intensivo durante más de 20 años, muestra que los HMM son lo suficientemente potentes como para modelar adecuadamente la mayor parte de las fuentes de variabilidad presentes en el habla, a pesar de la existencia de ciertos problemas bien conocidos: discriminación relativamente pobre, requerimientos explícitos en las asunciones sobre las distribuciones utilizadas, la no consideración de la correlación entre vectores acústicos, falta de adecuación en el modelado temporal, etc. **(F. Wessel, 1999)**

Dentro de la formulación genérica de los HMM, podemos hacer una clasificación en función de la naturaleza de las distribuciones que modelan las observaciones acústicas. Así, en una primera aproximación, podemos definir dichas distribuciones en un espacio discreto de símbolos, dando lugar a los Modelos discretos de Markov (DDHMM) **(Rabiner, 1991).** En este caso, las observaciones acústicas son símbolos pertenecientes a un alfabeto finito, con lo que se utilizan técnicas de cuantificación vectorial para transformar el vector de parámetros de entrada en uno de esos símbolos finitos.

Análogamente, podemos definir las distribuciones de probabilidad de las observaciones en un espacio continuo, dando lugar a los Modelos Continuos de Markov (CDHMM). En este caso, es necesario aplicar ciertas restricciones para limitar la complejidad de los procesos de estimación y cálculo de las probabilidades asociadas. El mecanismo más usual caracteriza cada modelo como una mezcla de funciones del mismo tipo generalmente gausianas.

Esta aproximación puede tener problemas de estimación fiable de sus parámetros y de demanda computacional, aunque se han desarrollado diversas técnicas para aliviarlos, basadas fundamentalmente en la reducción del número de parámetros a estimar generalmente compartiendo distribuciones o combinando modelos, en base a distintos criterios y en estrategias de optimización de cálculo.

Por último, y en la misma línea de reducir los inconvenientes que presentan tanto los DDHMM como los CDHMM, surgen los Modelos Semicontinuos de Markov (SCDHMM) **(Jesús Bernal Bermúdez),** en los que se comparte el mismo conjunto de funciones de densidad de probabilidad para distintos modelos, variando únicamente los pesos de ponderación aplicados a cada una de ellas, lo que puede verse como una unificación de los dos enfoques previos.

Comparándolo con la versión discreta, este modelado hace más exacto el proceso de cuantificación y robustece la estimación de probabilidad, al considerar varias distribuciones en cada caso, permitiendo además la estimación conjunta de los parámetros de la cuantificación y del modelo en sí. Comparándolo con la continua, reduce el número de distribuciones a considerar, robusteciendo la estimación en el

entrenamiento, aunque conservando la capacidad de modelado a partir de la mezcla de aquellas **(Good, 1953).**

Dado que los HMM no modelan con precisión la duración de las unidades de que se trate es posible incluir información sobre ella en el algoritmo. A lo largo del tiempo se han descrito en la literatura métodos para incluir explícitamente funciones de densidad de duración de estado en los HMM. Estudios previos que se relacionan en tratan las duraciones a nivel de modelo completo. Dicha información sobre las duraciones de las unidades se extraía durante los procesos de entrenamiento de los HMM, y se asumió que la distribución de la duración atribuida a una unidad es gauss.

## 2.2. Arquitecturas.

Cuando se enfrentan a la toma de decisiones en el proceso de diseño de un SRAH[17], una de las primeras preguntas a responder es la que atañe a la especificación de la arquitectura modular del sistema.

En la Figura 3 se muestra el esquema genérico de un SRAH, pero dada la generalidad del mismo, de cara a la toma de decisiones, se debe profundizar un poco en detalles de diseño y tratar de hacer una clasificación que abarque en lo posible elementos importantes en la arquitectura global. Esta clasificación afecta fundamentalmente al módulo de búsqueda, que es en el que se centra este apartado.

[17] **SRAH:** acrónimo de Sistema de Reconocimiento Automático de Habla

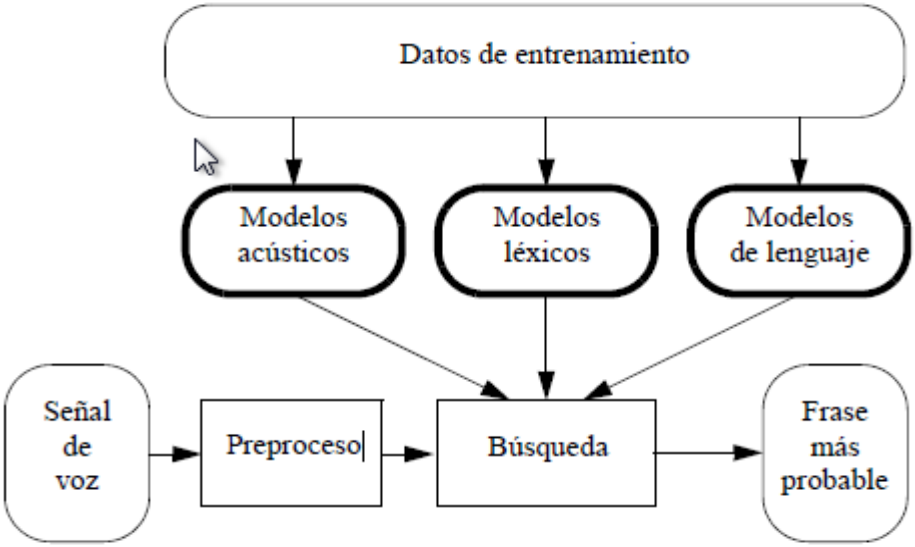



**Figura 3:** Esquema genérico de un Sistema de Reconocimiento Automático del Habla

En RAH, la señal de entrada se representa generalmente por una secuencia temporal de vectores de parámetros, que son calculados mediante análisis localizado. En el caso general, la idea final es transformar dicha representación inicial en un conjunto final de parámetros, más elaborados, en el sentido de estar más adaptados a la tarea de discriminación posterior.

Ahora sí, atendiendo al módulo de búsqueda acústica y tratando de buscar una perspectiva arquitectural general, se puede distinguir entre:

- Arquitecturas basadas en el paradigma *hipótesis-verificación*, en las que, conceptualmente, hay dos o más pasos de reconocimiento, es decir, se descompone la tarea en varios procesos en cascada, en las que se opera sobre un conjunto cada vez más reducido de hipótesis, utilizando modelos cada vez más potentes. En general, cada una de esas sucesivas etapas se caracteriza por:

    - Complejidad creciente: El coste computacional implicado para evaluar cada candidato es mayor a medida que avanzamos en la cadena de módulos de reconocimiento en cascada, debido a la mayor complejidad de los modelos y demás fuentes de información utilizadas.

- Resultados intermedios: Cada etapa ofrece a la siguiente un conjunto cada vez más reducido de alternativas entre las que decidir. Así, el incremento de complejidad del modelado se ve compensado por una menor demanda de cálculo al tener que operar sobre subconjuntos del espacio de alternativas inicial.

En estos sistemas cada etapa ofrece una hipótesis a la siguiente, que verificará dicha hipótesis, refinando el resultado, para proceder de la misma manera con la siguiente, o bien arrojará la decisión final, si es el último elemento de la cadena. Por supuesto debe asegurarse que la tasa obtenida por las etapas iniciales no limita significativamente la tasa final del sistema, y que la complejidad computacional final es menor que la que se obtendría con un sistema en un único paso.

Así, habrá que jugar con los parámetros: tamaño de la lista de hipótesis vs complejidad computacional de la etapa dada. En SRAH para gran y muy gran vocabulario, éste es el enfoque más usual **(SeMarket, 2010)**, utilizando para ello modelos de poca resolución acústica.

- Arquitecturas en las que, conceptualmente, es necesario un único proceso en bloque, desde la señal de voz hasta la obtención de la transcripción de la misma a la salida, a las que podríamos llamar basadas en el paradigma de verificación en un solo paso o, simplemente, verificación. Este enfoque implica, genéricamente, un mayor coste computacional, al tener que operar sobre el conjunto total de alternativas del espacio de búsqueda, pero suelen ofrecer mejores resultados finales y deben ser, obviamente, más robustos que los anteriores. Por supuesto, es posible introducir mecanismos de limitación del mismo, para reducir su coste, pero la idea central es el proceso en un único paso, sin que aparezcan listas de candidatos intermedias.

La clasificación vista no impone diferencias significativas en cuando a la algorítmica implicada, en el sentido de que cada algoritmo podrá utilizarse en un módulo de hipótesis o de verificación en función de su mayor o menos robustez. En este caso, y profundizando en una clasificación ortogonal a la anterior basándose en las distintas

aproximaciones a la hora de utilizar las fuentes de conocimiento disponibles, se puede hablar de:

- Sistemas no integrados, en los que cada uno de los módulos implicados utiliza parte de la información disponible para realizar la búsqueda. Un ejemplo de ello es el representado por los sistemas que extraen una cadena o malla de unidades elementales, seguidas por un módulo de acceso léxico que decide las hipótesis finales que con mayor probabilidad corresponderían a la cadena/malla de entrada. El módulo de generación de cadena o malla solamente utiliza información del modelado acústico, y es el de acceso léxico el que finalmente impone las restricciones extraídas de un diccionario.

- Sistemas integrados, en la que el resultado final (lista de preselección o candidato reconocido) se obtiene de forma directa, de modo que todas las fuentes de información disponibles se utilizan simultáneamente en el proceso de búsqueda. Este enfoque implica, de nuevo en general, mayor coste computacional que el anterior, ya que una división en varios módulos suele suponer una reducción de aquél; pero ofrecen mejores resultados al posibilitar la introducción de mecanismos de guiado de la búsqueda en el proceso desde el principio. Evidentemente, un sistema basado en hipótesis-verificación nunca podrá ser integrado ya que las fuentes de información se usan, forzosamente, de forma no simultánea.

Es esta última clasificación en la que se centra estudio, al ser la que impone las mayores restricciones en lo que a algorítmica se refiere, teniendo en cuenta que el enfoque integrado podrá utilizarse en un sistema basado en el paradigma de verificación siempre que las tasas de reconocimiento que consiga sean lo suficientemente altas para ser útiles en la tarea planteada. En caso contrario, siempre podrá utilizarse como módulo de generación de hipótesis en un sistema basado en hipótesis-verificación.

En lo que respecta a la algorítmica de la búsqueda acústica en sí, la aproximación más usual procede del campo del reconocimiento estadístico de patrones. En general, y dada la característica secuencial del proceso del habla, muchos de los problemas en SRAH pueden resolverse aplicando técnicas de programación dinámica, en las que se persigue

la determinación de un camino óptimo, o camino de mayor probabilidad (o menor coste), mediante la aplicación secuencial de optimizaciones (decisiones) locales, a lo largo de un espacio de búsqueda.

El primer gran impulso algorítmico en SRAH vino con el desarrollo y aplicación de las técnicas de alineamiento dinámico temporal (*Dynamic Time Warping*, DTW) en la que el objetivo es hacer una comparación entre un patrón de voz y una producción desconocida, de forma no lineal, permitiendo solucionar el problema difícilmente tratable hasta ese momento de la diferencia en longitud entre los patrones y las producciones (por la variabilidad temporal inherente al proceso del habla).

En la actualidad, es una técnica en desuso, por sus problemas de generalización, fundamentalmente. En su lugar, la misma base teórica de los algoritmos de programación dinámica, hace uso de la robustez del modelado estocástico (paramétrico) de los HMM para acometer tareas de todo tipo.

En este caso, para la parte acústica, se aplican las ideas propuestas sobre una base previa de sistemas de reconocimiento de habla conectada disponibles, utilizando la idea de construcción/reconocimiento de modelos de secuencias más complejas (frases o palabras) a partir de la concatenación de modelos de unidades más simples (alófonos, por ejemplo)**.**

Igualmente utilizaremos técnicas de acceso al léxico en los sistemas no integrados, básicamente fundamentados en los mismos principios de programación dinámica, pero aplicados a secuencias alfanuméricas, en lugar de vectores acústicos.

## 2.3. Modelos de lenguaje

La tarea de un modelo de lenguaje es capturar las restricciones que existen a la hora de combinar palabras para generar las frases posibles en un lenguaje dado. Dichas restricciones, de carácter sintáctico, semántico y pragmático, son difícilmente integrables como fuente de conocimiento en un SRAH. La importancia de dichos modelos es que permiten guiar de forma más eficaz a los reconocedores en el espacio de búsqueda sobre el que se mueven, si bien también pueden utilizarse para corregir a posteriori la salida de los mismos.

Sin embargo, en la literatura hay multitud de ejemplos de dicha integración, desde la utilización de diversos tipos de formalismos sintácticos y semánticos, hasta las gramáticas probabilísticas. En el primer caso, debido a problemas de cobertura y de complejidad computacional, su incorporación a los SRAH no ha sido ni mucho menos inmediata, y únicamente se han usado formalismos relativamente simples, basados en gramáticas regulares y de contexto libre, por ejemplo.

En sistemas de gran vocabulario, los métodos probabilísticos han sido los más utilizados, fundamentalmente por su adecuación al entrenamiento automático, utilizando grandes bases de datos etiquetadas convenientemente. Igualmente, el uso de técnicas específicas de suavizado les dotan de una gran robustez, habiéndose propuesto modelos basados en categorías, en lugar de palabras, perdiendo por tanto potencia en reducción de espacio de búsqueda, pero incrementando su generalidad y su facilidad de entrenamiento, aunque permanece sin solución definitiva el repertorio de categorías a utilizar.

Alternativas a esto lo constituyen mecanismos de estimación automática de dicho repertorio, aunque tampoco hay conclusiones claras sobre los criterios a utilizar, que generalmente se basan en disminución de la perplejidad, no habiéndose demostrado la relación directa entre este parámetro y la tasa de reconocimiento obtenida, dándose el caso de gramáticas con menor perplejidad que otras que, sin embargo, consiguen mejorar las tasas de reconocimiento finales. En toda esta discusión, hay que tener siempre presente el compromiso a establecer con la cobertura real que deseamos tener.

No se abundará más en este tema, por no ser objetivo de esta tesis profundizar en el mismo, de modo que se limitará al detallado estudio presentado en **(Jones, 1994)** La justificación de introducir aquí este tema es la necesidad de matizar cuantitativamente todos los resultados obtenidos en esta tesis aplicando el modelo de lenguaje adecuado si se plantea su uso en tareas en las que se procese habla continua o, incluso cuando se trate de habla aislada.

## 2.4. Técnicas de entrenamiento

Los sistemas de entrenamiento utilizados en SRAH tienen una importancia capital, al ser la base de estimación de los parámetros de los modelos acústicos y lingüísticos utilizados, modelos que se usarán en el proceso de reconocimiento.

Tradicionalmente, y centrándonos ya en reconocedores basados en HMM, el criterio más extendido para la estimación de estos ha sido el de máxima verosimilitud, y es ahí donde radica uno de los principales defectos de esa formulación: El criterio de máxima verosimilitud trabaja en el sentido de optimizar los parámetros de una distribución determinada en función de unos datos disponibles, pero el rendimiento de un SRAH se mide normalmente por su tasa de reconocimiento estimada. Así, no hay conexión directa entre el criterio de estimación usado para generar los HMM y la función objetivo final que queremos maximizar.

Un efecto lateral de este planteamiento es la ausencia de criterios de discriminación en el proceso de entrenamiento, con lo que los modelos no contienen en sí dicha propiedad, y es ahí donde, como se comentó anteriormente, las redes neuronales muestran su potencia. En este estado de cosas, durante los últimos años, se han venido desarrollando una serie de ideas orientadas a solucionar este problema y dotar de capacidad discriminadora explícita a los HMM.

En lo que respecta a la relación con las arquitecturas, en la literatura hay intentos de aplicar técnicas de entrenamiento dependiente y conjunto, en aquellos sistemas multi-módulo disponibles, en los que, en general, se hace independientemente.

## Conclusiones parciales

En este capítulo se realizaron un conjunto de investigaciones relacionadas con las principales técnicas que se utilizan en el mundo de hoy en el reconocimiento de habla. Las mismas han dado pie a tomar una decisión en cuanto a cuáles de estas tecnologías son más eficaces a la hora de realizar un RAH. Entre estos se vieron los HMM y las NNR que son de las tecnologías que más se utilizan en estos momentos para llevar a cabo la autenticación por voz.

# CAPÍTULO 3: Resultados Obtenidos.

## Introducción

En este capítulo se caracterizan y describen los algoritmos que componen la solución propuesta, se desarrolla un análisis crítico de cada una de las tendencias actuales en el desarrollo de sistemas de autenticación por habla, así como la fundamentación de su selección

## 3.1. Verificación del locutor

Partiendo de que el reconocimiento de locutores es la rama de la inteligencia artificial dedicada al desarrollo de reconocedores automáticos de locutores. Se diferencia del reconocimiento del habla en que, mientras que éste convierte un texto hablado a texto escrito, el reconocimiento de locutores decide quién de entre una lista de personas ha pronunciado el texto sobre el que se efectúa el reconocimiento, implicando una alta tasa de complejidad de los algoritmos a utilizar, esto se pueden clasificar en dos grandes grupos:

- Verificación del locutor dependiente del texto
- Verificación del locutor independiente del texto

El reconocimiento del locutor dependiente del texto está caracterizado por sesiones cortas de entrenamiento y sesiones de prueba. Las sesiones de entrenamiento consisten en varias repeticiones del léxico de entrenamiento, donde el habla total obtenida contiene generalmente de 4 a 8 segundos y los silencios son eliminados. **(Ivis Rodés Alfonso, 2009)**

Uno de los factores que influyeron en la decisión de utilizar la verificación del locutor dependiente del texto, fue teniendo en cuenta principalmente que los LVCSR[18] han sido usados en tareas independientes del texto, donde se reconocen de una conversación,

[18] **LVCSR** acrónimo de Sistemas de Reconocimiento del Habla para Vocabularios Grandes

las palabras más frecuentes dichas por cada locutor y a estas se le aplican técnicas del reconocimiento del locutor dependiente del texto. Esto implica una mejora en los sistemas de reconocimiento del locutor porque se restringen a un grupo específico de unidades. Pero, a su vez, aumenta el costo computacional de ese sistema por la existencia del LVCSR, que debe procesar largas conversaciones.

En la verificación del locutor dependiente del texto se distinguen dos métodos fundamentales:

- DTW[19]
- HMM[20]

El DTW consiste en comparar la locución de entrada con un conjunto de plantillas que representan las expresiones a reconocer. El entrenamiento se basa en almacenar en plantillas las expresiones a reconocer. Esas plantillas son conjuntos de rasgos acústicos ordenados en el tiempo. Para el reconocimiento se debe alinear de manera óptima la secuencia de rasgos de entrada con el modelo de referencia previamente almacenado. Al concluir la comparación, la distancia acumulada entre las dos expresiones es la base de la puntuación.

Este método es bastante simple y no requiere muchos recursos computacionales en la fase de entrenamiento. Se puede aplicar en sistemas de control de acceso con contraseña, teniendo previamente las plantillas de todas las posibles contraseñas. Esta es una desventaja de este método pues, al depender de las expresiones de referencia, imposibilita la variabilidad en la señal de voz. **(R. Cole, 1995)**

La técnica de modelado estadístico Modelos Ocultos de Markov ha sido muy utilizada en los campos de reconocimiento del habla y reconocimiento del locutor dependiente del texto, por su habilidad de modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz. Este método ha mostrado ser muy efectivo en el modelado y reconocimiento de fonemas, palabras y frases.

---

[19] **DTW** acrónimo de Alineamiento Dinámico en el Tiempo

[20] **HMM** acrónimo de Modelos Ocultos de Markov.

Las contraseñas, que consisten en secuencias de palabras tales como los dígitos, se utilizan mucho. En ellas, cada palabra está caracterizada por un HMM con un pequeño número de estados, donde cada estado es representado por una densidad de mezcla gausiana. Los parámetros del HMM son entrenados tomando varias repeticiones de la contraseña. De este proceso se obtiene el modelo de la contraseña y con los rasgos de la frase a verificar, se calcula la puntuación que permite decidir si aceptar o rechazar al cliente.

Esta técnica tiene una fundamentación estadística sólida con algoritmos de aprendizaje muy eficientes y además, posee una gran adaptabilidad a la variabilidad de las condiciones de la voz o del canal de transmisión. Dado que los HMM tienen menos cantidad de estados que ventanas en cada expresión, son mejores y más rápidos que los sistemas basados en DTW. Sin embargo, necesitan de muchos datos para el entrenamiento en aras de lograr una buena estimación de los parámetros del modelo.

## 3.2. Métodos de verificación utilizando modelos acústicos del habla

Actualmente, existen varios algoritmos de adaptación de los modelos del locutor, los más utilizados hasta la fecha son:

### 3.2.1. Reconocimiento de habla continua de gran vocabulario

En un sistema basado en GMM[21]-UBM[22] para la verificación de locutores restringidos en texto, se utilizan generalmente segmentaciones de palabras producidas por un sistema LVCSR[23], permitiendo al sistema enfocarse en las diferencias de locutores dentro de un conjunto de palabras.

El habla es segmentada en palabras y los verificadores GMM-UBM son entrenados y probados usando solo la voz de ese grupo de palabras. Para ello se usa un segmentador, generalmente un reconocedor del habla LVCSR, para dividir la voz de

[21] **GMM** acrónimo de Gaussian Mixture Models
[22] **UBM** acrónimo de Universal Background Model
[23] **LVCSR** acrónimo de Large Vocabulary Continuous Speech Recognition

entrada en palabras. La voz correspondiente a cada palabra es usada para entrenar el sistema GMM-UBM restringido a esas palabras. El UBM se entrena con un gran número de locutores, usando solamente la voz proveniente de ese grupo específico de palabras. **(Ivis Rodés Alfonso, 2009)**

La ventaja de este método viene de restringir el habla y comparar los mismos grupos de palabras pronunciadas por diferentes locutores. Las desventajas están dadas por:

- La necesidad de tener un buen segmentador, teniendo en cuenta que si su efectividad es baja eliminaría la especificidad de los modelos condicionados a las palabras.
- Requiere de grandes cantidades de información para el entrenamiento y la verificación, así como transcripciones del habla de alta calidad.

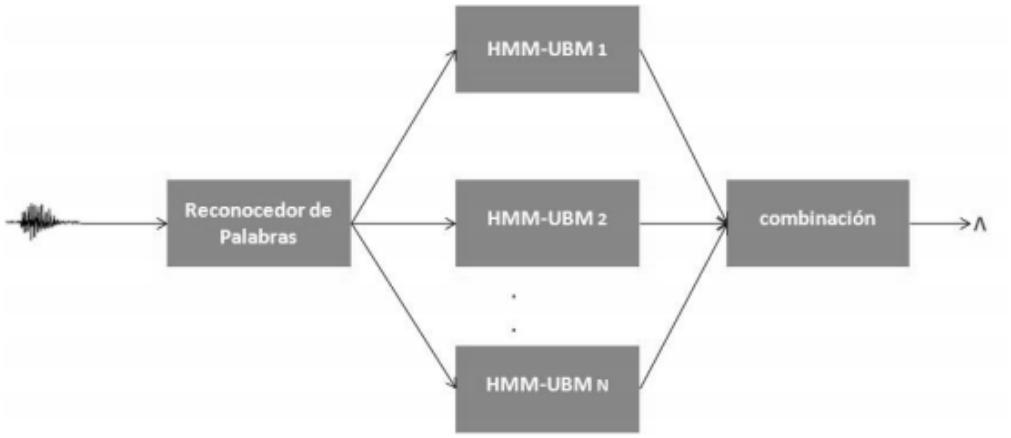



**Figura 4:** Sistema de Modelos Ocultos de Markoz combinado con *Universal Background Model* de verificación de locutor restringido en texto.

Este método también restringe el habla y se basa en el reconocimiento previo de palabras, necesitando también de un buen segmentador, lo que es una desventaja. Además, para modelar se usan modelos HMM, que poseen un alto costo computacional.

### 3.2.2. Modelos Ocultos de Markoz adaptados al locutor *Maximum a Posteriori*

La adaptación MAP[24] es un proceso de estimación en dos pasos. En el primer paso se estiman los estadísticos de los datos de entrenamiento para cada mezcla del UBM[25]. En

[24] **MAP** acrónimo de Maximum a Posteriori

el segundo paso, se combinan estos nuevos estadísticos con los estadísticos de los parámetros del UBM.

Dado un UBM y un vector de entrenamiento perteneciente al locutor que se desea Adaptar X = { x1, x2, x3… xT }, es necesario, primero, determinar el alineamiento probabilístico que existe entre el vector de entrenamiento y las mezclas que componen el UBM. Esto es, para la mezcla i del UBM, se calcula:

$$P(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}$$

A partir de ese término y de xT, se calculan los estadísticos necesarios para calcular a su vez los pesos, medias y varianzas:

$$n_i = \sum_{t=1}^{T} P(i|x_t)$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|x_t) x_t$$

$$E_i\left\{(x-\mu)^2\right\} = \frac{1}{n_i} \sum_{t=1}^{T} P(i|x_t) x_t^2$$

Finalmente, con estos nuevos estadísticos calculados de los datos de entrenamiento, se actualizan los estadísticos antiguos del UBM para cada mezcla i para obtener los parámetros adaptados:

$$\hat{w}_i = \left[\alpha_i^w n_i / T + \left(1-\alpha_i^w\right) w_i\right] \gamma$$

$$\hat{\mu}_i = \alpha_i^m E_i(x) + \left(1-\alpha_i^m\right) \mu_i$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + \left(1-\alpha_i^v\right)\left(\hat{\sigma}_i^2 + \hat{\mu}_i^2\right) - \hat{\mu}_i^2$$

[25] **UBM** acrónimo de Universal Background Model

$\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ son los coeficientes que controlan el balance entre las estimaciones antiguas y nuevas de los pesos, medias y varianzas, respectivamente. Estos coeficientes se definen como sigue:

$$\alpha_i^p = \frac{n_i}{n_i + r^p}, p \in \{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$$

siendo $r^p$ un factor fijo de relevancia para el parámetro P. Además, $\gamma$ es un factor de escala que se calcula sobre todos los pesos adaptados para asegurar que éstos suman uno. La adaptación MAP, al combinar los modelos GMM de los locutores con los UBM, ha sido clave en la mejora de los clasificadores.

Uno de los problemas a los que se enfrenta esta técnica es a los pocos datos de entrenamiento que existe en el reconocimiento dependiente del texto. La adaptación MAP ayuda a solucionar este problema introduciendo el modelo UBM. Sin embargo para diferentes locutores, la topología del modelo HMM tiende a ser la misma debido a la forma de adaptación.

Por ello, en este trabajo se proponen dos métodos no supervisados: UBM local y UBM global, capaces de obtener la topología de un HMM para cada locutor y así aumentar la capacidad discriminativa de cada modelo, lo que hace más robusta la verificación. Los resultados experimentales mostraron que ambos son efectivos para trabajar con pocas observaciones. **(Ivis Rodés Alfonso, 2009)**

El modelo UBM local trabaja muy bien bajo diferentes condiciones de entrenamiento. El modelo UBM global puede superar al HMM adaptado si los datos del entrenamiento tienen un volumen moderado. Una vez obtenida la topología del HMM por cualquiera de estos métodos, se realiza la adaptación MAP para refinar los parámetros del modelo.

### 3.2.3. Modelos Ocultos de Markoz adaptados al locutor *Maximum Likelihood Linear Regression*

La adaptación MLLR[26] consiste en transformar matrices de medias y opcionalmente matrices de covarianza de un modelo HMM mediante una transformación afín que maximice la función de similaridad dados los nuevos datos de adaptación y el modelo:

$$\bar{\mu} = A\mu + b$$
$$\bar{\sum} = H \sum H^T$$

donde μ es el vector de medias en el modelo, ∑ es la matriz de covarianza, μ y ∑ las matrices de media y covarianza adaptadas respectivamente, (A, b) es la transformación afín para la adaptación de la media y H la matriz de transformación para la adaptación de la covarianza. Para encontrar los parámetros óptimos, se usa generalmente el método de maximización de la expectancia en dos pasos: estimar la transformación de la media, dados A y b y luego estimar la transformación H de la covarianza.

La adaptación MLLR es una técnica especialmente desarrollada para la adaptación de modelos HMM independientes del locutor, a la voz de un locutor en particular a partir de un número limitado de expresiones y por lo tanto, se puede aplicar al problema de reconocimiento del locutor dependiente de texto. Además, en los casos en los que se cuenta con pocas expresiones para realizar la adaptación, la adaptación MLLR consigue mejores resultados si se agrupan en clases y se transforman las medias de toda la clase, utilizando para ello una misma transformación lineal. **(Ivis Rodés Alfonso, 2009)**

De esta manera, la adaptación MLLR reduce el número de parámetros a entrenar, pasando de depender linealmente del número de gausianas a depender linealmente del número de clases. Esto la convierte en una técnica muy robusta de adaptación de modelos HMM a un locutor, incluso cuando se utilizan modelos más complejos.

El sistema de reconocimiento del habla realiza una primera descodificación usando los coeficientes MFCC y un modelo del lenguaje bi-grama. Las hipótesis resultantes son usadas para adaptar un segundo conjunto de modelos basados en rasgos PLP. Estos modelos adaptados son usados en un segundo paso de descodificación que está restringido por tri-gramas que generan las listas de los N-mejores. Estos son recalculados

[26] **MLLR** acrónimo de Maximum Likelihood Linear Regression

por un modelo del lenguaje cuatri-grama y por modelos prosódicos, hasta llegar a la palabra final.

La transformada MLLR se aplica en los dos pasos del reconocimiento. En el primero, se basa en un modelo de fonemas de referencia, usando tres transformadas: para no-voz, para fonemas sonoros y no sonoros. El segundo paso de descodificación está basado en las palabras de referencia generadas por el primer paso, y se aplican nueve tipos de transformadas diferentes a las clases: no voz, vocales altas/bajas, consonantes sonoras, explosivas sonoras/no sonoras, fricativas sonoras/no sonoras y nasales. Los coeficientes de una o más transformadas de adaptación son concatenados en un vector de rasgos y modelados usando SVM. **(Ivis Rodés Alfonso, 2009)**

El SVM es entrenado para cada locutor usando los rasgos de un conjunto de entrenamiento de background como muestras negativas y los datos de los locutores como muestras positivas. Además, el rango dinámico del vector de coeficientes es normalizado, que reemplaza cada valor del rasgo por su rango en la distribución de background, esta normalización realiza un reescalado adaptado de los rasgos para obtener una distribución aproximadamente uniforme.

El método propuesto está compuesto por los siguientes componentes: extracción de vectores de rasgos a partir de muestras de habla de locutores por medio de una etapa de adaptación de un reconocedor del habla y construcción de una función discriminante del locutor mediante SVM. Este método de adaptación de rasgos ha dado muy buenos resultados, estando a la altura e incluso superando a los métodos cepstrales y fue evaluado como el mejor rasgo en un estudio del estado del arte de los rasgos más utilizados para el reconocimiento del locutor.

El método CMLLR aplicado a los sistemas de reconocimiento del locutor permite extraer rasgos que están más enfocados a las características relacionadas con el locutor. El método presentado posee dos etapas. En la primera, se construye un modelo universal de background GMM/UBM a partir de los rasgos cepstrales del locutor. En la segunda, se estiman las transformadas CMLLR para cada locutor de interés usando el UBM creado, obteniéndose un vector de rasgos de alta dimensión por locutor, el cual se modela usando las SVM.

La ventaja de esta técnica sobre la propuesta por es que el proceso de entrenamiento no depende de la trascripción ni del lenguaje, y así captura las diferencias entre los rasgos acústicos independientes del locutor y dependientes del locutor. Sin embargo, dado que se usa un modelo GMM para estimar la transformada CMLLR, esta es menos precisa y probablemente más dependiente del mensaje. Este método combinado con sistemas MFCC-SVM y MFCC-GMM tiene rendimientos muy significativos.[27]

## 3.3. Arquitectura

Basándose en el estudio realizado en el capítulo anterior se propone como arquitectura a utilizar en el desarrollo de sistemas de autenticación por reconocimiento de habla, la basada en el paradigma *hipótesis-verificación.* Este es un elemento muy importante, teniendo en cuenta la creciente complejidad computacional que implican los sistemas de software de este tipo, será muy provechoso el empleo de la programación de sistemas distribuidos, así como el uso de clúster de procesamiento, precisamente este tipo de arquitectura propuesta propicia la utilización de los mismos.

Partiendo de que un sistema distribuido se define como una colección de computadoras separados físicamente y conectados entre sí por una red de comunicaciones distribuida; cada máquina posee sus componentes de hardware y software que el usuario percibe como un solo sistema. El usuario accede a los recursos remotos (RPC) de la misma manera en que accede a recursos locales, o un grupo de computadores que usan un software para conseguir un objetivo en común.

Por otra parte el término clúster se aplica a los conjuntos o conglomerados de computadoras construidos mediante la utilización de componentes de hardware comunes y que se comportan como si fuesen una única computadora. El cómputo con clusters surge como resultado de la convergencia de varias tendencias actuales que incluyen la disponibilidad de microprocesadores económicos de alto rendimiento y redes de alta velocidad, el desarrollo de herramientas de software para cómputo distribuido de alto

**[27] Para más información sobre los HMM referirse a los Anexos.**

rendimiento, así como la creciente necesidad de potencia computacional para aplicaciones que la requieran.

Otro elemento a tener en cuenta es que las aplicaciones paralelas escalables requieren: buen rendimiento, baja latencia, comunicaciones que dispongan de gran ancho de banda, redes escalables y acceso rápido a archivos. Un clúster puede satisfacer estos requerimientos usando los recursos que tiene asociados a él.

Los *clusters* ofrecen las siguientes características a un costo relativamente bajo:

- Alto rendimiento
- Alta disponibilidad
- Alta eficiencia
- Escalabilidad

La tecnología clúster permite a las organizaciones incrementar su capacidad de procesamiento usando tecnología estándar, tanto en componentes de hardware como de software que pueden adquirirse a un costo relativamente bajo.

## 3.4. Metodología de Desarrollo del Software

Actualmente con el creciente desarrollo tecnológico y la aparición de nuevos modelos de producción, han ido apareciendo nuevas metodologías de proceso de desarrollo. Estos procesos han ido tomado características marcadas, es por ello que cuando se habla de los paradigmas de desarrollo se hace referencia a la filosofía o enfoque de desarrollo para una determinada metodología, estas pueden ser agrupadas en:

- Metodologías Orientadas al Plan
- Metodologías Ágiles

Las metodologías orientadas al plan, también conocidas como metodologías tradicionales o clásicas, son aquellas que están guiadas por una fuerte planificación durante todo el proceso de desarrollo, donde se realiza una intensa etapa de análisis y diseño antes de

la construcción del sistema. Ejemplo de metodologías que entran en esta clasificación entra RUP y METRICA 3. **(Larman, 2004)**

Se entiende como desarrollo ágil de software a un paradigma de desarrollo de software basado en procesos ágiles. Los procesos ágiles de desarrollo de software, conocidos anteriormente como metodologías livianas, intentan evitar los tortuosos y burocráticos caminos de las metodologías tradicionales enfocándose en la gente y los resultados. Las metodologías más utilizadas dentro de esta clasificación se encuentran: Programación Extrema (XP), Scrum, Feature Driven Development (FDD), Adaptive Software Development (ASD), RUP Ágil, existiendo dos variantes AUP y EUP por tan solo citar algunas. **(sanabria, 2008)**

Lo primero que se debe tener en cuenta para su selección es el alcance y envergadura de la solución propuesta. Partiendo de que el objetivo de esta investigación es desarrollar un sistema de autenticación por reconocimiento del habla y que no es tarea fácil, se necesitaría una metodología que se adapte a proyectos grandes y que permita controlar de manera transparente todo el proceso de desarrollo, fundamentalmente producirlo en el tiempo y costo esperado, debido a la necesidad de un sistema como este para el país.

RUP es considerada como una metodología sumamente adaptable al contexto y necesidades de cada organización que tiene como objetivo asegurar la producción de software con calidad, dentro de plazos y presupuestos predecibles. Esto lo hace extensible para proyectos de cualquier envergadura, a diferencia de otras metodologías famosas como XP o FDD, orientadas a proyectos pequeños y de corta duración.

El uso de esta permite aumentar la productividad de los desarrolladores mediante acceso a base de conocimiento, plantillas y herramientas. Se centra en la producción y mantenimiento de modelos del sistema más que en producir documentos.

## 3.5. Lenguaje de Modelado

Todo proyecto de software requiere de etapas de modelado que permitan experimentar y visualizar el sistema que desea construir. El modelado de un sistema de software consiste en representar un conjunto de objetos que tienen un significado ingenieril, ajustándose a un conjunto de normas y signos de representación determinados por la

notación o lenguaje de modelado seleccionado, precisando de esta forma los modelos de mayor importancia para el entendimiento de las partes y el todo de la aplicación, incluyendo su entorno.

La falta de estandarización en la manera de representar gráficamente un modelo impedía que los diseños gráficos realizados se pudieran compartir fácilmente entre distintos desarrolladores. En consecuencia a esto, Jacobson, Booch y Rumbaugh deciden crear el UML. Concebido como un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema de software, incluyendo aspectos conceptuales tales como procesos de negocios y funciones del sistema y aspectos concretos como expresiones de lenguajes de programación, esquemas de bases de datos y componentes de software reutilizables.

Es importante destacar que UML no es un proceso de desarrollo. No describe los pasos sistemáticos a seguir para desarrollar software, sólo permite documentar y especificar los elementos creados mediante un lenguaje común describiendo modelos.

Para modelar un sistema complejo no es suficiente una única representación, sino que se requieren múltiples modelos, donde cada uno representa una vista o aspecto del sistema en concreto, que pueden ser representados con diferentes grados de precisión y a su vez se deben complementar entre sí. Es por ello que los elementos de UML se muestran mediante diagramas que presentan múltiples vistas del sistema, ese conjunto de vistas son conocidos como modelos.

RUP propone UML como lenguaje de modelado para representar todos los esquemas de un sistema de software, de hecho sus desarrolladores fueron los mismos. Partiendo de que un enfoque sistemático permite construir estos modelos de una forma consistente demostrando su utilidad en sistemas de gran envergadura como la solución propuesta, se puede afirmar que RUP y UML es la combinación perfecta para desarrollarla.

## 3.6. Tecnologías

En sentido general un paquete de librerías o biblioteca no es más que una lista de instrucciones bien definidas, ordenadas y de carácter finito, que permite hallar la solución a un determinado problema. Estas proporcionan un conjunto de servicios a programas

independientes, permitiendo que el código y los datos se compartan, para que de esta forma puedan modificarse modularmente.

Uno de los criterios más importantes a tener en cuenta para seleccionar el lenguaje de programación es el nivel de abstracción, el cual se refiere a cuan cercano es o no, al idioma humano, para ello identificamos tres clasificaciones.

Los lenguajes de bajo nivel son aquellos que proporcionan poca o ninguna abstracción del microprocesador de un ordenador. En esta clasificación se encuentran como lenguaje de primera generación el código máquina y posteriormente el lenguaje ensamblador. Teniendo en cuenta que los circuitos micro programables son sistemas digitales, lo que significa que trabajan con dos únicos niveles de tensión, dichos niveles por abstracción, se simbolizan con el 0 y el 1, es por ello que el lenguaje de máquina sólo utiliza dichos signos, por tanto no pueden ser escritos o leídos usando un editor de texto y es raro que una persona lo use directamente.

Contrario a esto el lenguaje ensamblador es un tipo de lenguaje de bajo nivel utilizado para escribir programas informáticos ofreciendo una mayor capa de abstracción y consiste en una serie de instrucciones que corresponden al flujo de órdenes ejecutables que pueden ser cargadas en la memoria de una computadora.

Es importante tener en cuenta que cada arquitectura de computadoras tiene su propio lenguaje de máquina y en consecuencia su propio lenguaje ensamblador. Los ordenadores difieren en el tipo y número de operaciones que soportan, también pueden tener diferente cantidad de registros y distinta representación de los tipos de datos en memoria. Aunque la mayoría de las computadoras son capaces de cumplir esencialmente las mismas funciones, la forma en que lo hacen difiere y los respectivos lenguajes ensambladores reflejan tal diferencia.

El lenguaje ensamblador fue utilizado ampliamente para el desarrollo de software, pero actualmente sólo se utiliza en contadas ocasiones, especialmente cuando se requiere la manipulación directa del hardware o se pretenden rendimientos inusuales de los equipos,

debido que al programar en él se trabajan con los registros de memoria de la computadora de forma directa.

Los lenguajes de alto nivel se caracterizan por expresar los algoritmos de una manera adecuada a la capacidad cognitiva humana, en lugar de a la capacidad ejecutora de las máquinas, son independientes de la arquitectura del ordenador, permitiéndole al programador abstraerse por completo del funcionamiento interno de la máquina para la que está diseñando el programa. Ejemplo de lenguajes dentro de esta clasificación se encuentran C++, Java, C#, Python, etc.

Los lenguajes de medio nivel se encuentran en un punto intermedio, pues se apropian de los elementos más importantes de las otras clasificaciones. Dentro de estos podría situarse C, este incorpora muchos elementos propios del ensamblador, puede acceder a los registros del sistema, trabajar con direcciones de memoria, con la particularidad de que podemos realizar las operaciones mucho más legibles, utilizar estructuras de datos y otras características propias de los lenguajes de alto nivel. **(Coplien, Diciembre 1994)**

Son precisos para ciertas aplicaciones como la creación de sistemas operativos, permitiendo un manejo abstracto independiente del hardware, a diferencia del ensamblador, pero sin perder mucho del poder y eficiencia que tienen los lenguajes de bajo nivel.

Teniendo en cuenta que los procesadores usados en las computadoras solo son capaces de entender y actuar según lo indican programas escritos en lenguaje de máquina, podemos clasificar los demás según su ejecución en:

- Lenguajes Compilados.
- Lenguajes Interpretados.
- Lenguajes basados en Máquinas Virtuales.

Un programa que se escribe en un lenguaje de alto nivel tiene que traducirse a un código que pueda ser interpretado por la máquina. Una vez escrito, se traduce a partir de su código fuente por medio de un compilador, en un archivo ejecutable para una determinada plataforma. Entiéndase por compilador aquella aplicación encargada de

traducir un lenguaje de alto nivel al código máquina para una determinada arquitectura, depositándolo en un fichero binario, que un sistema operativo será capaz de cargar en la memoria principal y pedirle al CPU que lo ejecute. Tal es el caso de lenguajes como C, C++, Fortran, Pascal, Delphi, etc.

Contrario a esto los lenguajes de programación interpretados fueron diseñados para ser ejecutados por medio de un intérprete, también se les conoce como lenguajes de Script. Entiéndase por intérprete aquella aplicación encargada de analizar y ejecutar programas escritos en un lenguaje de alto nivel, estos sólo realizan la traducción a medida que sea necesario, típicamente instrucción por instrucción y por lo general no guardan el resultado de dicho proceso. Entre los lenguajes que caen en esta clasificación podemos mencionar a ActionScript, JavaScript, ASP, PHP, Perl, Python, Ruby, etc. **(Millán, 1998)**

Los lenguajes basados en máquinas virtuales como bien se había explicado anteriormente toman elementos de ambas clasificaciones, pues realizan una compilación, pero su resultado no sería entonces código máquina, sino una especie de código intermedio que solo se ejecuta a través de su propia plataforma de software. En esta clasificación se sitúan lenguajes como Java y C#.

Para dar cumplimiento a los objetivos propuestos de este trabajo y teniendo en cuenta la complejidad en cuestiones de procesamiento o coste computacional de los algoritmos propuestos, se necesitaba un lenguaje de programación lo más cercano posible al código de máquina, pero que al mismo tiempo no sea tan complicado como Ensamblador o C, este tendría que ser capaz de proveer los recursos de los lenguajes de alto nivel, como la programación orientada a objeto. En consecuencia a esto se definió C++ como el lenguaje de programación que más se ajusta a las necesidades de esta investigación.

El C++ es un lenguaje de propósito general basado en el C, diseñado a mediados de los años 1980, por Bjarne Stroustrup, a diferencia de su antecesor este ha añadido nuevos tipos de datos, clases, plantillas, funciones virtuales, mecanismo de excepciones, sistema de espacios de nombres, funciones online, sobrecarga de operadores, referencias, sobrecarga de funciones, operadores para manejo de memoria persistente y algunas utilidades adicionales basadas en librerías externas.

La creciente complejidad de los sistemas modernos, ha provocado cada vez más, que aquellos objetos que anteriormente solo necesitaban algunos métodos para definir sus funcionalidades, crezcan en complejidad hasta el punto de tornarse inmanejables, por la excesiva cantidad de instrucciones que pueden llegar contener.

Este lenguaje provee de muchos recursos que permiten dar solución a estos problemas, tal es el caso de la abstracción de datos y la programación genérica, conceptos que ponen el trabajo desempeñado por las computadoras más cerca del punto de vista humano, permitiéndole que se adapte a múltiples situaciones. **(Bronson, 2000)**

Por otra parte la fuente que provee la flexibilidad de los lenguajes de programación son los punteros, si se analiza a profundidad el tema, tanto C# como Java los tienen implícito, solo que estos implementan un conjunto de capas de abstracción de forma tal que son transparente para el desarrollador. El uso de puntadores permite realizar operaciones de asignación dinámica de memoria y manipular estructuras de datos dinámicas, es por ello que tienen una fuerte relación con el manejo eficiente de tablas y estructuras más complejas.

Desde este punto de vista tanto la asignación como liberación de memoria dinámica es responsabilidad del programador, estos tienen que ser capaces de construir elegantes mecanismos y jerarquías de clases, que controlen correctamente la creación y destrucción de objetos. Claro está que un conocimiento mediocre o incompleto realmente impedirá desarrollar programas eficientes, pero se debe tener bien claro que el C++ fue diseñado por y para programadores, un criterio diferente no sería suficientemente profesional.

No se quiere decir con esto que se desechen los lenguajes de mayor grado de abstracción y se comience a programar en ensamblador o código máquina, sino que se valore realmente desde un punto de vista ingenieril y sepan identificar el potencial que este ofrece, sin dejarse guiar por la comodidad o el afán de ganar dinero fácil, todo lo contrario, piensen en desarrollar sus aplicaciones lo más robustas y eficientes posibles.

Como bien se había expresado anteriormente, trabajar con formatos de audio es extremadamente complicado a nivel de programador, en consecuencia a esto surgieron

interfaces que abstraen la complejidad y diversidad de las primitivas del hardware, entre las que se encuentran:

- Spro[28] versión 4.0
- Alize.

Spro es un conjunto de herramientas desarrolladas por Guillaume Gravier para el procesamiento de señales de voz. Estas herramientas proveen comandos que implementan algoritmos de extracción de características estándares para aplicaciones de reconocimiento de voz y de locutor.

Por otra parte ALIZE es una plataforma de software que tiene el objetivo de facilitar el desarrollo de aplicaciones en el área de reconocimiento de voz y del locutor. Es una biblioteca desarrollada en C++ en el LIA[29] de la Universidad de Avignon en Francia, por Frederic Wils bajo la dirección de Jean Francois Bonastre, en febrero del año 2003. **(LIA, 2008)**

Alize está compuesta de dos niveles distintos:

- Nivel base, donde se encapsula la complejidad técnica de los módulos (adquisición de datos, cálculo, almacenamiento, etc.). Este nivel evita que el usuario tenga que administrar la memoria directamente.
- En un segundo nivel se incluyen las utilidades y algoritmos que se manipulan por el usuario (listas de administración, inicialización de modelos, algoritmos MAP, etc.).

Alize presenta una documentación muy bien detallada para su uso, también tiene las siguientes características:

- Tiene un nivel de funcionamiento que corresponde con el estado del arte actual, en términos de error pero también en términos de recursos de cómputo necesarios.
- Facilita el desarrollo de demostraciones y aplicaciones prácticas.

---

28 **Spro** acrónimo de Speech Signal and Processing Toolkit

29 **LIA** acrónimo de Laboratorio de Informática de Avignon

Se propone utilizar las herramientas de la biblioteca SPro para extraer 19 coeficientes MFCC, más la energía, así como los parámetros delta y doble delta, con lo cual se obtuvo un vector de dimensión 60. Para las etapas de detección de energía, normalización, entrenamiento de los modelos y cálculo de la puntuación se propone de la biblioteca LIA_RAL/ALIZE.

Para el enfoque discriminativo basado en Máquinas de Vectores de Soporte se propone el *kernel* de secuencia discriminante lineal generalizada o GLDS. Los vectores de características son los mismos utilizados en el enfoque generativo. La expansión polinomial fue calculada con el empleo de la biblioteca LIA_RAL/ALIZE. Se propone la biblioteca LibSVM para las etapas de entrenamiento y predicción de las SVM. **(Addison-Wesley, Abril 2001)**

Este lenguaje provee de muchos recursos que permiten dar solución a estos problemas, tal es el caso de la abstracción de datos y la programación genérica, conceptos que ponen el trabajo desempeñado por las computadoras más cerca del punto de vista humano, permitiéndole que se adapte a múltiples situaciones.

Por otra parte si bien es cierto que las aplicaciones que se desarrollen con lenguajes como C++ no son portables para otras plataformas, por el simple hecho de que son compilados para una arquitectura en específico. Se asegura que sí es posible desarrollar aplicaciones para diferentes arquitecturas con un mismo código fuente.

Al igual que se desarrollan aplicaciones en lenguajes como java, aunque estos únicamente requieren de una máquina virtual específica para cada plataforma, manteniendo el mismo estándar de codificación, en C++ es conocido como “ISO C++ o ANSI C++” cuyas normas que rigen los estándares del lenguaje y todo compilador que se respete de adoptarlas, por tanto solo se necesita del compilador específico para la plataforma que está destinada el producto final, con la diferencia que estos generan código nativo con un alto grado de optimización en memoria y velocidad, lo que lo convierte en uno de los lenguajes más eficientes. **(Coplien, Diciembre 1994)**

## 3.1. Validación de la solución propuesta

La solución propuesta se basa principalmente para su validación en un estudio minucioso de una serie de artículos desarrollados por profesionales de una elevada categoría científica, así como años de experiencia en temas afines a esta investigación, los cuales pertenecen CENATAV[30].

El CENATAV es un centro orientado a las investigaciones teóricas y aplicadas en el área del Reconocimiento de Patrones y la Minería de Datos. Las investigaciones incluyen en la actualidad el procesamiento digital de imágenes y señales, la teledetección, el reconocimiento lógico combinatorio de patrones, el reconocimiento sintáctico estructural, la teoría de testores, algoritmos conceptuales, el análisis de texturas, la interpretación conceptual de datos espaciales, la minería de texto, la minería de datos mezclados, entre otras. Las aplicaciones están dirigidas a áreas tales como la biometría, la recuperación de información, la prospección geológica, procesamiento de información de texto, entre otras. **(CENATAV, 2004)**

Otro elemento importante a tener en cuenta es que este centro presta diversos servicios científicos y técnicos tales como consultoría e información científica especializada en las esferas del Reconocimiento de Patrones y de la Minería de Datos. Además, divulga los resultados de las investigaciones y participa activamente en eventos científicos nacionales e internacionales, promoviendo el desarrollo de vínculos con instituciones y organizaciones afines, siendo esto de vital importancia para la UCI, en aras de lograr un convenio de trabajo que le permitiría desarrollar un auténtico sistema RHA, contando con un equipo multidisciplinario y de experiencia.

Entre los artículos más significativos que tuvieron en cuenta para determinar el modelo teórico de la solución propuesta, específicamente la selección de emplear los modelos ocultos de Markoz adaptados al locutor MLLR se encuentran:

- Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor.

[30] **CENATAV**: Centro de Aplicaciones de Tecnologías de Avanzada

- Reconocimiento del locutor dependiente del texto con modelos acústicos del habla.
- Autenticación biométrica por el habla del usuario en las redes de info-comunicaciones.
- Noise robust voice detector for speaker recognition.
- Autenticación Biométrica por el Habla del Usuario en las Redes de Info-comunicaciones.

Todos estos artículos han sido certificados y algunos han sido publicados fuera del ámbito nacional, también se consultaron otras bibliografías pertenecientes diferentes autores fuera y dentro del ámbito nacional, las cuales se encuentran reflejadas al final del documento específicamente en los acápites de Bibliografías y Referencias Bibliográficas.

## Conclusiones parciales

Teniendo en cuenta lo que se plantea en los anteriores capítulos y el estudio que fue realizado para poder así definir cuál sería el modelo arquitectónico que se debe utilizar para construir un sistema de RHA, se ha llegado a la elaboración del capítulo que acaba usted de leer en el cual se exponen las principales ideas de Metodologías, Modelos Arquitectónicos y Lenguajes de programación que se creyó fueran los más indicados para desarrollar un sistema de este tipo.

Entre estas tecnologías fueron seleccionadas las que más pudieran aportar y las que más facilidades y seguridad ofrecieran para el desarrollo del sistema. Para el mismo se propuso como lenguaje de programación el C++ debido a un conjunto de funcionalidades y estándares que el mismo proporciona para el tratamiento de sonidos y utilizando algoritmos probabilísticos como los HMM se pueden obtener con la vinculación de los dos, un buen resultado a la hora del desarrollo.

# CONCLUSIONES

La verificación del locutor dependiente del texto con rasgos acústicos es un tema que ha recibido mucha atención en el campo del reconocimiento del locutor, por la variedad de aplicaciones que posee. En este trabajo se realizó un estudio exhaustivo de la literatura relacionada con esta temática, y se estableció una clasificación de los métodos más utilizados.

De ellos, los que mejores resultados han obtenido son los que modelan al locutor mediante HMM adaptados al mismo, ya sea por adaptación MAP o MLLR. Resultados obtenidos que se expusieron en este trabajo ponen en evidencia la superioridad de la adaptación MLLR sobre MAP, en tareas dependientes del texto, por su capacidad de adaptar el HMM a un locutor con pocos datos.

En este campo se ha investigado bastante y se han propuesto muchos métodos con el objetivo de lograr una verificación del locutor precisa, pero aún existen problemas abiertos que no tienen la solución más óptima, por ejemplo: los grandes volúmenes de datos a procesar, la complejidad computacional de los algoritmos de entrenamiento y prueba , las diferencias en el léxico, en los datos de entrenamiento y los de la prueba, los pocos datos disponibles para el entrenamiento, las diferencias en los canales de entrenamiento y prueba y la detección de tramas con voz en presencia de ruido.

Por lo tanto, queda mucho que hacer para lograr un sistema capaz de reconocer una persona por su voz, que sea tan competente como el cerebro de una persona.

# RECOMENDACIONES

Con el objetivo de socializar el conocimiento acumulado durante todo este proceso investigativo y de afianzar estos elementos cognoscitivos que son tan complejos e importantes, se recomienda llevar a la práctica cada uno de los elementos aquí expuestos, así como realizar un estudio más profundo dirigido a la combinación de los HMM con Redes Neuronales, en aras de ganar en cuanto a fiabilidad, permitiendo de esta forma optimizar todo el poseso de autenticación basado en el rasgo biométrico más conocido por voz.

Por otra parte en función de completar o extender el estudio realizado y teniendo en cuenta que los clústeres son usualmente empleados para mejorar tanto el rendimiento, como la disponibilidad por encima de la que es provista por un solo computador típicamente siendo más económico que computadores individuales de rapidez y disponibilidad comparables, se recomienda que se desarrolle un estudio más profundo en la configuración y puesta en práctica de esta tecnología, aplicada a sistemas de autenticación por reconocimiento del habla.

## Trabajos citados


- Addison-Wesley. Abril 2001. ***The C++ Programming Language.*** **s.l. : Prentice Hall, Abril 2001. 0-201-88954-4.**

- Adriana Becerra, Marcela Gómez, María Fernanda Ordóñez, Byron Macas, Ing. Janeth Chicaiza. August 3, 2009. **Reconocimiento de voz mediante Modelos Ocultos de Markov. August 3, 2009.**

- Aquino, Orlando Fernández. 2004. ***El Aparato Fonador.*** **Uberlandia, Brasil : s.n., 2004.**

- Basso, Gustavo. 1999. ***Análisis Espectral. La transformada de Fourier en la Música.*** **Argentina : UNLP, 1999. 950-3401-50-X.**

- Boix, Joaquim Llisterri. 1991. ***Introducción a la fonética: el método experimental.*** **Barcelona : Anthropos, 1991. 9788476583029 .**

- Bourlard, N. Morgan y H. May 1995. ***Continuous Speech REcognition: An introduction to the hybrid HMM/connectionist Approach.*** **s.l. : EEE Signal Processing magazine, May 1995. volume 12, number 3, pages 24-42.**

- Bronson. 2000. ***C++ Para Ingeniería Y Ciencias.*** **s.l. : Thomson , 2000. 9687529873 .**

- Calvo, José R., Fernández, Rafael y Hernández, Gabriel. 2008. **Autenticación Biométrica por el Habla del Usuario en las Redes de Info-comunicaciones. Habana : Memorias de Segurmática 2008, 2008.**

- Calvo, José Ramón, y otros. 2009. ***Autenticación biométrica por el habla del usuario en las redes de info-comunicaciones.*** **Habana : CD Memorias de Informática 2009, 2009. 978-959-286-010-0.**

- Cambridge University Engineering. 2005. ***The HTK Book (for HTK Version 3.4).*** **2005.**

- CENATAV. 2004. **Centro de Aplicaciones de Tecnologías de Avanzada. [En línea] 2004. [Citado el: 7 de Mayo de 2010.] http://www.cenatav.co.cu/es/index.html.**

- Cole, Ronald A. 1998. ***Survey of the State of the Art of Human Language Technology.*** **s.l. : Cambridge University Press, 1998. 978-0521592772.**

- Coplien, James O. Diciembre 1994. ***Advanced C++ Programming Styles and Idioms.*** **s.l. : Prentice Hall, Diciembre 1994. 0-201-54855-0.**

- Davis, Gary y Jone, Ralph. 1990. ***The Sound Reinforcement Handbook.*** **Milwukee, USA : Hal Leonard Publishig, 1990.**

- Espinosa Duró, Virginia. 2004. ***Evaluación de Sistemas de Reconocimiento Biométrico.*** **Barcelona : s.n., 2004.**

- F. Wessel, R. Schuter, K. Macherey, and Hermann Ney. 1999. **Confidence measures for large vocabulary continuous speech recognition.** ***In IEEE, Transactions on Speech and Audio Processing.*** **1999. Vols. volume 9, pages 288–298.**

- Federico, Moises Castellanos y Jesus. 2007. ***Reconocimiento de Voz con Redes Neurales.*** **s.l. : Universidad Simon Bolıvar, 2007.**

- ***Feel the Noise.*** Boulware, Jack. Octubre de 1999. **San Francisco, California : Condé Nast Publications, Octubre de 1999.**

- Fernández, Rafael, Calvo, José R. y Hernández, Gabrie. 2008. **Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor.** ***Serie Azul.*** **Habana : Reporte Técnico CENATAV, 2008. Vol. 008.**

- Fletcher, Harvey. 1995. ***Speech and Hearing in Communication.*** **Woodbury, USA : American Institute of Physics, 1995.**

- Frega, Ana Lucía, Fernández Calvo, Diana y Ratto, Jorge. 2000. ***Sonido, música y ecoacústica : dimensiones educativas del fenómeno sonoro.*** **Buenos Aires : s.n., 2000. 950-503-309-5.**

- Geomodeling Technology Corp. 2009. **Geomodeling. [En línea] 2009. [Citado el: 12 de Enero de 2010.] www.geomodeling.com.**

- Gish, M. Siu and H. 1999. **Evaluation of word confidence for speech recognition systems. 1999. Vols. 13, pages 299–318.**

- Gonzalez, Mike. 1985. ***Collins Concise Spanish-Engish English-Spanish Dictionary.*** **New York, USA : Prentice Hall Press, 1985.**

- Good, I. J. 1953. **The population frequencies of species and the estimation of population parameters. 1953. Vols. pages 237–264.**

- Hernando Pericás, Francisco Javier. mayo 1993. **Técnicas de procesado y representación de la señal de voz para el reconocimiento del hala en ambientes ruidosos. Barcelona : s.n., mayo 1993.**

- Ivan Jacobson, Grandy Boovh y Jame Rumbuagh. 2004. ***El Proceso Unificado de Desarrollo del Software.*** **La Habana : Félix Varela, 2004.**

- Ivis Rodés Alfonso, Dr. C.José Ramón Calvo de Lara. 2009. **Reconocimiento del locutor dependiente del texto con modelos acústicos de habla. Habana : Serie Azul, 2009. Vol. 2148, 2072-6287.**

- Jacob Benesty, M. Mohan Sondhi, Yiteng Huang. 2008. ***Springer Handbook of Speech Processing.*** **s.l. : Springer-Verlag Berlin Heidelberg, 2008.**

- Jesús Bernal Bermúdez, Jesús Bobadilla Sancho y Pedro Gómez Vilda. 2000. ***Reconocimiento de Voz y Fonética Acústica.*** **s.l. : Alfaomega, Ra-Ma , 2000. 970-15-0541-7.**

- John R. Deller. John H. L. Hansen, John G. ***Discrete-Time Processing of Speech Signals.***

- Jones, G. 1994. **Application of Linguistic Models to Continuous Speech Recognition. s.l. : Universidad de Bristol, 1994.**

- Labrada, Jerónimo. 1995. ***El registro sonoro.*** **Santafé de Bogotá : s.n., 1995. 958-02-1005-5.**

- Larman, Craig. 2004. ***UML y Patrones.*** **La Habana : Felix Varela, 2004.**

- Lawrence Rabiner, Biing Hwang Juang. ***Fundamentals of Speech Recognition.*** **s.l. : Prentice-Hall International.**

- LIA. 2008. **Mistral.** ***Mistral.*** **[En línea] Laboratoire Informatique of Avignon, 2008. [Citado el: 14 de 4 de 2010.] http://mistral.univ-avignon.fr/en/.**

- Llorach, E. Alarcos. 1986. **Fonología Española. Madrid : s.n., 1986.**

- Martínez Celdrán, Eugenio. 2007 . ***Análisis espectrográfico de los sonidos del habla.*** **Barcelona : Ariel, 2007 . 978-84-344-8271-5.**

- Millán, José Antonio Jiménez. 1998. ***Compiladores y Procesadores de Lenguaje.*** **s.l. : Thomson , 1998. 84-96274-39-X.**

- Miyara, Federico. 1999. ***Acústica y Sistemas de Sonido.*** **Rosario : Editorial UNR Editora, 1999.**

- —. 2006. ***Acústica y Systemas de Sonido.*** **Rosario : UNR Editora, 2006.**

- —. 2000. ***Control de Ruido.*** **Rosario : ASOLOFAL, 2000.**

- —. 2010 . **La Voz Humana.** ***Escuela de Ingeniería Electrónica .*** **[En línea] Febrero de 2010 . http://www.eie.fceia.unr.edu.ar/~acustica/biblio/fonatori.pdf.**

- —. 2009. **Potencia.** ***Escuela de Ingeniería Electrónica.*** **[En línea] Febrero de 2009. http://www.eie.fceia.unr.edu.ar/~acustica/biblio/potencia.pdf.**

- Mujica, José. 2008. **Ingeniería de Audio.** ***Escuela Superior de Audio.*** **[En línea] 2008. [Citado el: 7 de Mayo de 2010.] http://www.escuelasuperiordeaudio.com.ve/Ampca/INGENIERIA_DE_AUDIO.htm.**

- Nuance Communications. 2010. **Nuance. *Nuance.* [En línea] 2010. [Citado el: 3 de Febrero de 2010.] http://www.nuance.com/.**

- Pelton, Gordon E. 1992. ***Voice Processng.*** **Singapore : MacGraw-Hill, 1992.**

- Pressman, Roger S. 2005. ***Ingenieria de Software un Enfoque Práctico.*** **La Habana : Felix Varela, 2005.**

- Puig, Sergi Jordà. 1997. **Audio digital y MIDI, Guías Monográficas Anaya Multimedia. Madrid : s.n., 1997.**

- Quatieri, Thomas F. 2002. ***Discrete-Time Speech Signal Processing: Principles and Practice.*** **s.l. : Prentice Hall, 2002. 0-13-242942-X.**

- R. Cole, ed. J. Mariani, H. Uszkoreit, A. Zaenen, y V. Zue. 1995. **Survey of the State of the Art in Human Language Technology. [En línea] 1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.**

- Rabiner, B. H Juang and L. R. 1991. **Hidden Markov Models for speech recognition. 1991. Vol. 33, 3.**
- Rayleigh, J. W. S. 1894. ***The Theory of Sound.*** **New York, United States : Dover, 1894.**

- Reccasens i Vives, Daniel. 1993. ***Fonètica i Fonologia.*** **[Enciclopèdia Catalana] Barcelona, España : s.n., 1993.**

- Richard O. Duda, PeterE. Hart,David G.Stork, Wiley-Interscience. 2000. ***Pattern Classsification.*** **s.l. : Wiley-Interscience, 2000. 978-0471056690.**

- Rochaix, Edmundo Carlos y Carlos, Juan Garay. 2010. **Auditorias técnicas en obras en ejecución mediante la investigación de los ruidos y vibraciones. *SOCIEDAD ESPAÑOLA DE ACÚSTICA.* [En línea] 2010. http://www.sea-acustica.es/publicaciones/4350jh038.pdf.**

- Rodés, Ivis y Calvo, José R. 2009. ***Reconocimiento del locutor dependiente del texto con modelos acústicos del habla.*** **Habana : Serie Azul, 2009. Vol. 009. 2072-6287.**

- sanabria, willian. 2008. **Métodos Ágiles. *Métodos Ágiles.* [En línea] 22 de 09 de 2008. [Citado el: 15 de 04 de 2009.]**

- Sanchis, Claudio F. Estienne y Alberto. 2006. ***Sistema de Reconocimiento Automatico de Habla basado en Maxima Entropıa.*** **Buenos Aires, Argentina : s.n., 2006.**

- Schapper, Dr. Paul y Rivolta, Dr. Mercedes. Diciembre 2004. **Autenticación & Firmas Digitales en E- Legislación y Seguridad. [En línea] Diciembre 2004. http://www.mdbegp.org/www/LinkClick.aspx?fileticket=t6CjQF5zpgQ%3D&tabid=66&mid=386&language=es-ES.**

- SeMarket. 2010. **SeMarket. [En línea] 2010. [Citado el: 3 de Febrero de 2010.] http://www.semarket.com/es/soluciones/control.acceso/biocloser.php.**

- Silva, Dr. José Luis Batista. 2005. **Mapping Interactivo. [En línea] Noviembre de 2005. http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=1051.**

- Simón, Pablo Iglesias. 2004. **El diseñador de sonido: función y esquema de trabajo. [En línea] Agosto de 2004. http://www.pabloiglesiassimon.com/textos/El%20disenador%20de%20sonido.pdf. 1133-8792.**

- Smith, Steven W. 1997. ***The Scientist and Engineer's Guide to Digital Signal Processing.*** **California Technical : s.n., 1997.**

- Stevens, K. Ñ. 1998. ***Acoustic Phonetics.*** **s.l. : MIT Press, 1998.**

- Stroustrup, Bjarne. Noviembre 1999. ***The Design and Evolution of C++.*** **s.l. : Prentice Hall, Noviembre 1999. 0-201-54330-3.**

- Syrdal, A., Bennet, R. y Greenspan, S. 1995. ***Applied Speech Technology.*** **FL, USA : CRC Press, 1995.**

- TAPIADOR MATEOS, MARINO y SIGÜENZA PIZARRO, JUAN ALBERTO. 2005. ***TECNOLOGÍAS BIOMÉTRICAS APLICADAS A LA SEGURIDAD.*** **s.l. : RA-MA EDITORIAL, 2005. 978-84-7897-636-2.**

- Tribaldos, Clemente. 1993. ***Sonido Profesional.*** **Madrid, España : Paraninfo, 1993.**

- Trubetzkoy, Nicolai S. 1992 . ***Principios de fonología.*** **Madrid : Cincel, 1992 . 84-7046-444-2.**

- Universidad de La Rioja. 2001-2010. **Dialnet. [En línea] 2001-2010. [Citado el: 15 de Enero de 2010.] http://dialnet.unirioja.es/servlet/articulo?codigo=1119628.**

- Wu Chou, Biing Hwang Juang. ***Pattern Recognition in Speech and Language Processing.*** **s.l. : Georgia Institute of Technology.**

- X.D. Huanga, H.W. Hona, M.Y. Hwanga and K.F. Leea. 2002. **A comparative study of discrete, semicontinuous, and continuous hidden Markov models. s.l. : Computer Speech & Language, 2002. Vol. 7, 4.**

- Xuendong, Huang; Fileno, Alleva; Hon, Hsiao-Wuen; Mei-Y. 1992. **The SPHINX-II Speech Recognition System: on Overview. s.l. : School of Computer Science, 1992.**

## Bibliografía


- Addison-Wesley. Abril 2001. ***The C++ Programming Language.*** **s.l. : Prentice Hall, Abril 2001. 0-201-88954-4.**

- Adriana Becerra, Marcela Gómez, María Fernanda Ordóñez, Byron Macas, Ing. Janeth Chicaiza. August 3, 2009. **Reconocimiento de voz mediante Modelos Ocultos de Markov. August 3, 2009.**

- Bourlard, N. Morgan y H. May 1995. ***Continuous Speech REcognition: An introduction to the hybrid HMM/connectionist Approach.*** **s.l. : EEE Signal Processing magazine, May 1995. volume 12, number 3, pages 24-42.**
- Bronson. 2000. ***C++ Para Ingeniería Y Ciencias.*** **s.l. : Thomson , 2000. 9687529873 .**

- Cambridge University Engineering. 2005. ***The HTK Book (for HTK Version 3.4).*** **2005.**

- Cole, Ronald A. ***Survey of the State of the Art of Human Language Technology.***

- Coplien, James O. Diciembre 1994. ***Advanced C++ Programming Styles and Idioms.*** **s.l. : Prentice Hall, Diciembre 1994. 0-201-54855-0.**

- F. Wessel, R. Schuter, K. Macherey, and Hermann Ney. 1999. **Confidence measures for large vocabulary continuous speech recognition.** ***In IEEE, Transactions on Speech and Audio Processing.*** **1999. Vols. volume 9, pages 288–298.**

- Federico, Moises Castellanos y Jesus. 2007. ***Reconocimiento de Voz con Redes Neurales.*** **s.l. : Universidad Simon Bolıvar, 2007.**

- Frega, Ana Lucía, Fernández Calvo, Diana y Ratto, Jorge. 2000. ***Sonido, música y ecoacústica : dimensiones educativas del fenómeno sonoro.*** **Buenos Aires : s.n., 2000. 950-503-309-5.**

- Geomodeling Technology Corp. 2009. **Geomodeling. [En línea] 2009. [Citado el: 12 de Enero de 2010.] www.geomodeling.com.**

- Gish, M. Siu and H. 1999. **Evaluation of word confidence for speech recognition systems. 1999. Vols. 13, pages 299–318.**

- Good, I. J. 1953. **The population frequencies of species and the estimation of population parameters. 1953. Vols. pages 237–264.**
- Hernando Pericás, Francisco Javier. mayo 1993. **Técnicas de procesado y representación de la señal de voz para el reconocimiento del hala en ambientes ruidosos. Barcelona : s.n., mayo 1993.**

- Ivan Jacobson, Grandy Boovh y Jame Rumbuagh. 2004. ***El Proceso Unificado de Desarrollo del Software.*** **La Habana : Félix Varela, 2004.**

- Ivis Rodés Alfonso, Dr. C.José Ramón Calvo de Lara. 2009. **Reconocimiento del locutor dependiente del texto con modelos acústicos de habla. s.l. : Serie Azul, 2009. Vol. 2148, 2072-6287.**

- Jacob Benesty, M. Mohan Sondhi, Yiteng Huang. 2008. ***Springer Handbook of Speech Processing.*** **s.l. : Springer-Verlag Berlin Heidelberg, 2008.**

- Jesús Bernal Bermúdez, Jesús Bobadilla Sancho y Pedro Gómez Vilda. ***Reconocimiento de Voz y Fonética Acústica.*** **s.l. : Ra-Maâ .**

- John R. Deller. John H. L. Hansen, John G. ***Discrete-Time Processing of Speech Signals.***

- Jones, G. 1994. **Application of Linguistic Models to Continuous Speech Recognition. s.l. : Universidad de Bristol, 1994.**

- Labrada, Jerónimo. 1995. ***El registro sonoro.*** **Santafé de Bogotá : s.n., 1995. 958-02-1005-5.**

- Larman, Craig. 2004. ***UML y Patrones.*** **La Habana : Felix Varela, 2004.**

- Lawrence Rabiner, Biing Hwang Juang. ***Fundamentals of Speech Recognition.*** **s.l. : Prentice-Hall International.**

- LIA. 2008. **Mistral.** ***Mistral.*** **[En línea] Laboratoire Informatique of Avignon, 2008. [Citado el: 14 de 4 de 2010.] http://mistral.univ-avignon.fr/en/.**

- Llorach, E. Alarcos. 1986. **Fonología Española. Madrid : s.n., 1986.**

- Millán, José Antonio Jiménez. 1998. ***Compiladores y Procesadores de Lenguaje.*** **s.l. : Thomson , 1998. 84-96274-39-X.**

- Nuance Communications. 2010. **Nuance.** ***Nuance.*** **[En línea] 2010. [Citado el: 3 de Febrero de 2010.] http://www.nuance.com/.**

- Pressman, Roger S. 2005. ***Ingenieria de Software un Enfoque Práctico.*** **La Habana : Felix Varela, 2005.**

- Puig, Sergi Jordà. 1997. **Audio digital y MIDI, Guías Monográficas Anaya Multimedia. Madrid : s.n., 1997.**

- Quatieri, Thomas F. 2002. ***Discrete-Time Speech Signal Processing: Principles and Practice.*** **s.l. : Prentice Hall, 2002. 0-13-242942-X.**

- R. Cole, ed. J. Mariani, H. Uszkoreit, A. Zaenen, y V. Zue. 1995. **Survey of the State of the Art in Human Language Technology. [En línea] 1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.**

- Rabiner, B. H Juang and L. R. 1991. **Hidden Markov Models for speech recognition. 1991. Vol. 33, 3.**

- Richard O. Duda, PeterE. Hart,David G.Stork, Wiley-Interscience. ***Pattern Classsification.***

- sanabria, willian. 2008. **Métodos Ágiles. *Métodos Ágiles.* [En línea] 22 de 09 de 2008. [Citado el: 15 de 04 de 2009.]**

- Sanchis, Claudio F. Estienne y Alberto. ***Sistema de Reconocimiento Automatico de Habla basado en Maxima Entropıa.* Buenos Aires, Argentina : s.n.**

- SeMarket. 2010. **SeMarket. [En línea] 2010. [Citado el: 3 de Febrero de 2010.] http://www.semarket.com/es/soluciones/control.acceso/biocloser.php.**

- Silva, Dr. José Luis Batista. 2005. **Mapping Interactivo. [En línea] Noviembre de 2005. http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=1051.**

- Smith, Steven W. 1997. ***The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical : s.n., 1997.**

- Stevens, K. Ñ. 1998. ***Acoustic Phonetics.* s.l. : MIT Press, 1998.**

- Stroustrup, Bjarne. Noviembre 1999. ***The Design and Evolution of C++.* s.l. : Prentice Hall, Noviembre 1999. 0-201-54330-3.**

- Universidad de La Rioja. 2001-2010. **Dialnet. [En línea] 2001-2010. [Citado el: 15 de Enero de 2010.] http://dialnet.unirioja.es/servlet/articulo?codigo=1119628.**

- Wu Chou, Biing Hwang Juang. ***Pattern Recognition in Speech and Language Processing.* s.l. : Georgia Institute of Technology.**

- X.D. Huanga, H.W. Hona, M.Y. Hwanga and K.F. Leea. 2002. **A comparative study of discrete, semicontinuous, and continuous hidden Markov models. s.l. : Computer Speech & Language, 2002. Vol. 7, 4.**

- Xuendong, Huang; Fileno, Alleva; Hon, Hsiao-Wuen; Mei-Y. 1992. **The SPHINX-II Speech Recognition System: on Overview. s.l. : School of Computer Science, 1992.**

## Anexo #1: Elementos de un modelo oculto de Markov discreto

- **N**: número de estados del modelo
  - estados, s = {s1 , s2 , . . . , sN }
  - estado en tiempo t, q t ∈ s
- **M**: número de símbolos de observación (ej., observaciones discretas)
  - símbolos de observación, v = {v 1 , v2 , . . . , vM }
  - observación en tiempo t, o t ∈ v
- **A** = {aij }: distribución de la probabilidad de la transición del estado
  - aij = P(qt+1 = sj |qt = si ), 1 ≤ i, j ≤ N
- **B** = {bj (k)}: distribución de la probabilidad del símbolo de observación del estado j
  - bj (k) = P(vk at t|qt = sj ), 1 ≤ j ≤ N, 1 ≤ k ≤ M
- **π** = {πi }: distribución del estado inicial
  - πi = P(q1 = si ), 1 ≤ i ≤ N

Desde una perspectiva notación, un HMM se escribe típicamente como:
λ = {A, B, π}

## Anexo #2: Tres problemas básicos de HMM

1. **Puntuación**: Dada una secuencia de observación O = {o1 , o2 , ..., oT } y un modelo λ = {A, B, π}, ¿cómo calculamos P(O | λ), la probabilidad de la secuencia de observación?

**==>** Algoritmo de avance-retroceso

2. **Ajuste**: Dada una secuencia de observación O = {o 1 , o2 , ..., oT }, ¿cómo elegimos una secuencia de estado Q = {q 1 , q2 , ..., qT } que de algún modo sea óptima?

**==>** Algoritmo de Viterbi

3. **Entrenamiento**: ¿Cómo ajustamos los parámetros del modelo λ = {A, B, π} para maximizar P(O | λ)?

**==>** Procedimientos de re estimación de Baum-Welch

## GLOSARIO

- **Timbre**: Es la capacidad que nos permite diferenciar los sonidos.
- **Potencia acústica**: El nivel de potencia acústica es la cantidad de energía radiada en forma de ondas por unidad de tiempo por una fuente determinada.
- **La fonología**: es un subcampo de la lingüística. Mientras que la fonética estudia la naturaleza acústica y fisiológica de los sonidos o alófonos, la fonología describe el modo en que los sonidos funcionan (en una lengua en particular o en las lenguas en general) en un nivel abstracto o mental.
- En fonética, se llama **alófono** a cada uno de los fonos o sonidos que en un idioma dado se reconoce como un determinado fonema, sin que las variaciones entre ellos tengan valor diferenciativo; cada fono corresponde a una determinada forma acústica, pero en las reglas de la lengua se los considera como poseyendo el mismo valor.
- **Inteligencia Artificial:** la rama de la ciencia informática dedicada al desarrollo de agentes racionales no vivos.
- **Parametrizar:** en idioma inglés se conoce como "to costumize", consiste en adecuar algo (darle parámetros) para que se ajuste dinamicamente atraves de factores externos.
- **Perceptrón:** es un tipo de red neuronal artificial desarrollado por Frank Rosenblatt, también puede entenderse como perceptrón la neurona artificial y unidad básica de inferencia en forma de discriminador lineal, que constituye este modelo de red neuronal artificial, esto debido a que el perceptrón puede usarse como neurona dentro de un perceptrón más grande u otro tipo de red neuronal artificial.
- **Vectores equiespaciados:** conjunto de magnitudes físicas separadas por valores equidistantes, que tienen en cuenta la dirección y el sentido.