Universidad de las Ciencias Informáticas

Facultad # 2

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.

"Propuesta del algoritmo de reconocimiento biométrico de locutores sobre IP para la plataforma PlaTel"

Autores: Nayansi Castellón Otero.

Yaretmy Hernández Moreno.

Tutora: Ing. Angélica M. Díaz Valdivia.

Ciudad de La Habana, julio del 2010.

"Año 52 de La Revolución"

“Educar a un joven no es hacerle aprender algo que no sabía, sino hacer de él alguien que no existía”.

John Ruskin.

Declaración de autoría

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de ____del año____.

Nayansi Castellón Otero

_____

Autor

Yaretmy Hernandez Moreno

_____

Autor

Ing. Angélica M. Díaz Valdivia

_____

Tutor

Dedicatoria:

A mi mamita: Por todo su cariño, su paciencia, por confiar en mí, por cuidarme y apoyarme en todo momento, por dedicarme cada segundo de su vida y demostrarme que soy lo más importante. Por secar mis lágrimas y disfrutar mis momentos de alegría, por ser mi amiga, mi guía, mi fuerza, por ayudarme a realizar este sueño.

Te quiero mamita.

Nayansi Castellón Otero.

A todas aquellas personas que han sabido estar a mi lado en los momentos que más lo he necesitado. A toda mi familia, pero en especial a ese tesoro que la vida me dio el privilegio de tener como madre y a mi papito lindo que a pesar de no haber vivido todo este tiempo junto a mí, lo quiero con todo mi corazón. Sin ustedes este sueño hoy no sería una realidad. La eternidad no sería suficiente para retribuirles todo el amor y el cariño que me han brindado, todos los esfuerzos y sacrificios que han hecho para que hoy yo pueda cumplir mis deseos. Para ustedes este inmenso regalo. Los amo.

Yaretmy Hernandez Moreno.

Agradecimientos:

A mi mamá por darme las fuerzas para seguir adelante.

A mis abuelos por ser tan lindos conmigo.

A mi esposo por estar a mi lado en todo momento, por su cariño y comprensión, por dejar de vivir su vida para simplemente vivir la mía.

A mi compañera de tesis por compartir conmigo esta dura tarea.

A mis grandes amigos los 101 dálmatas por darme el placer de compartir con ellos tantos momentos hermosos.

A nuestra tutora la Ing. Angélica M. Díaz Valdivia por ayudarnos, apoyarnos y guiarnos.

Al Ing. Duany Baró Menéndez por su ayuda incondicional.

A mis nuevas y mis viejas amistades por soportarme siempre y darme la oportunidad de ser su amiga.

A todas las personas que han contribuido al desarrollo de este trabajo y en mi formación como profesional.

Al Comandante en Jefe por darme la oportunidad de ser parte de esta gran obra que es la Universidad de las Ciencias Informáticas que un día abrió sus puertas para convertirme en la persona que hoy soy.

Nayansi Castellón Otero.

Agradecimientos

Agradezco a mi mamí por ser esa madre que me dio tanto cariño y que entregó todo su amor y ternura incondicionalmente, a ti por ser tan especial y estar siempre junto a mí apoyándome en los duros momentos, aguantando mis malcriadeces con paciencia. Por ser esa personita que ha puesto todo su empeño en criarme de la mejor forma posible y que ha respetado todas las decisiones que he tomado en mi vida. Por haber luchado con garras, engrandeciéndote en momentos de desesperación y de angustia. A ti mi madre querida, agradeceré eternamente todo lo que logré hacer en este mundo. Gracias por existir.

A mi papi, por ser un padre luchador, comprensivo, que supo darse cuenta en el momento preciso que lo necesitaba y que le quería mucho. A ti padre, por haberme ayudado tanto, por darme tanto cariño en los momentos que podemos estar juntos, que aunque sean cortos, para mí siempre serán eternos. A ti por ser un ejemplo para mí, por quererme y guiarme en la vida, por ser ese padre que tanto admiro y quiero, por tus esfuerzos y deseos de verme realizada. A ti por ser ese talismán que adorna mi vida con felicidad, amor y cariño, siempre te estaré agradecida. Gracias por estar ahí para mí.

A mis hermanas, por todo ese amor incondicional que me han dado desde que era una niña, por ayudarme en todo momento, por estar atentas a todo lo que me sucede, por preocuparse y apoyarme. A ustedes mis queridas hermanitas, mi mayor gratitud. Las palabras no son suficientes para expresar todo el amor que les tengo. Mi mayor deseo es retribuirles todo lo que me han sabido ofrecer desde niña. Si un día volviera a nacer y tuviera la oportunidad de elegir a mis hermanas, sin duda alguna, ustedes ocuparían ese lugar tan especial. Gracias por ser tan buenas. Las quiero.

Agradecimientos

A toda mi familia porque de un modo u otro, todos han puesto su granito de arena contribuyendo en mi formación como profesional. Feliz y agradecida de tener este hermoso regalo, ustedes.

A Jorge e Ismary por haber sido mi segunda familia durante 4 años, tratándome con amor y cariño, ayudándome en aquellos momentos tristes y difíciles que pasé. Les agradezco con todo mi corazón toda la ayuda que me supieron dar desde el primer día que me acogieron en la familia con tanto afecto. Gracias.

A mis compañeros de aula por haber pasado momentos tan lindos junto a ellos.

A Javier Ernesto por ser ese chico creativo y juguetón que me ha hecho reír, convirtiéndose en un ser especial y único. Gracias por darme tu cariño y hacerme tan feliz en tan poco tiempo.

A nuestra tutora Ing. Angélica M. Díaz Valdivia, por ayudarnos y guiarnos en la realización de este trabajo.

Al Ing. Duany Baró Menéndez por apoyarnos en el momento preciso.

A todas las personas que han contribuido en mi formación como profesional y en la realización del presente trabajo.

Yaretmy Hernandez Moreno.

RESUMEN

El reconocimiento del locutor es otro ejemplo de identificación biométrica personal que ha sido aplicado en diversas áreas, despertando el interés de múltiples investigadores y desarrolladores de todo el mundo.

En la Universidad de las Ciencias Informáticas, se encuentra en desarrollo la plataforma de telecomunicaciones unificada PlaTel. Esta plataforma tiene como objetivo principal establecer comunicación por: correo, SMS, fax, jabber, y VoIP, para ello utiliza programas de software importantes como: Openfire, Hylafax, Postfix y PBX IP Asterisk respectivamente. Esta PBX es una central telefónica con capacidad para VoIP, que brinda diversos servicios, entre los que se encuentran: correo de voz, conferencia, matutino, respuesta de voz iterativa (IVR), colas de atención, identificador de llamante, música y llamada en espera, reportación de llamada, entre otros. Uno de los aspectos necesarios que se desea llevar a cabo en PlaTel, es lograr la calidad de los servicios que se brindan (QoS) y la seguridad de todos los elementos que se manejan en la red. Aunque se han obtenido algunos resultados en los puntos anteriormente mencionados, en servicios como el correo de voz e IVR, la PBX no tiene incorporado un sistema de reconocimiento que posibilite la identificación del locutor, por lo que el objetivo del presente trabajo de investigación es proponer el algoritmo de reconocimiento biométrico de locutor más adecuado a la plataforma PlaTel.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN.

El desarrollo de las comunicaciones y el avance tecnológico, han hecho inminente la interacción hombre-máquina, como un nuevo medio que proporciona, comodidad, rapidez y practicidad a las actividades que se realizan en el entorno cotidiano, mejorando así la calidad de vida. Son muchas las acciones que se ejecutan hoy mediante un ordenador, dígase transferencia de información, controles remotos, acceso a datos, etc., por lo que es necesario establecer normas que garanticen la seguridad de la información que está siendo empleada en cualquier ámbito.

En muchos casos, los métodos de autenticación biométrica han demostrado ser una buena opción, ya que se basan en características biométricas tales como: iris, huellas digitales, estructuras genéticas, entre otras, rasgos que no se pueden perder ni olvidar, además de ser difíciles de imitar. La voz, es una característica biométrica de gran aceptabilidad, es por ello que el reconocimiento del locutor, ha despertado el interés de múltiples investigadores y desarrolladores de todo el mundo, convirtiéndose en otro ejemplo de identificación biométrica personal.

Este reconocimiento incluye la identificación del locutor, esta técnica, tiene la misión de identificar a quién pertenece una voz desconocida, o sea, le asigna a la persona, la identidad del individuo registrado que mejor se aproxime a las características de la señal de voz. Por otra parte se encuentra la verificación, la cual deberá decidir si la persona que declara una cierta identidad es o no quien dice ser.

Para llevar a cabo estas prácticas se hace necesario incluir bases de datos que almacenen archivos de sonido con grabaciones de palabras de los clientes, algo que se consigue en las sesiones de entrenamiento donde los usuarios pronunciarán varias frases, mediante las cuales estos sistemas puedan hacer posteriormente las comparaciones necesarias. Ante estas acciones el sistema puede responder de tal forma que acepte al usuario registrado o rechace al intruso, teniendo en cuenta los resultados obtenidos.

Para garantizar un buen funcionamiento existen diferentes algoritmos de reconocimiento como: Modelos Ocultos de Markov, Redes Neuronales, Cuantización Vectorial, Modelo de Mezclas Gaussianas, Máquinas de Vectores de Soporte, entre otros, los cuales han sido desarrollados para determinar un nivel de coincidencia entre las pronunciaciones, comparando para ello, las locuciones de entrada con las locuciones que genere el usuario en cada intento de acceso, de modo que se obtenga un reconocimiento

eficaz. En general, estos métodos son los encargados de cuantificar, extraer y clasificar las características de la voz de cada locutor, existiendo algunos que lo realicen con mayor desempeño, debido a su construcción y sus características internas.

Todos los aspectos anteriormente mencionados se han venido utilizando en múltiples esferas, entre las que se puede mencionar fundamentalmente la telefonía IP, servicio que permite realizar llamadas telefónicas sobre redes IP u otras redes de paquetes utilizando un PC o un teléfono, especialmente diseñado para soportar este tipo de comunicaciones. Esto puede realizarse entre dos usuarios de PC conectados a internet, un usuario conectado puede además llamar a un teléfono fijo y de este último puede llamarse a una central que esté conectada a internet y pedir que se le comunique con otro teléfono. Es un servicio muy usado que ofrece múltiples ventajas, básicamente lo que se hace es digitalizar las voces y comprimirlas en paquetes de datos hasta llegar al destino, donde se convierte nuevamente en voz.

En la Universidad de las Ciencias Informáticas, se encuentra en desarrollo la plataforma de telecomunicaciones unificada PlaTel. Esta plataforma tiene como objetivo principal establecer comunicación por: correo, SMS, fax, jabber, y VoIP, para ello utiliza programas de software importantes como: Openfire, Hylafax, Postfix y PBX IP Asterisk respectivamente. Esta PBX es una central telefónica con capacidad para VoIP que brinda diversos servicios, entre los que se encuentran: correo de voz, conferencia, matutino, respuesta de voz iterativa (IVR), colas de atención, identificador de llamante, música y llamada en espera, reportación de llamada, entre otros. Uno de los aspectos necesarios que se desea llevar a cabo en PlaTel, es lograr la calidad de los servicios que se brindan (QoS) y la seguridad de todos los elementos que se manejan en la red. Aunque se han obtenido algunos resultados en los puntos anteriormente mencionados, en servicios como el correo de voz e IVR, la PBX no tiene incorporado un sistema de reconocimiento que posibilite la identificación del locutor, por lo que surge como **problema científico:**

¿Cómo lograr un adecuado reconocimiento del locutor en la plataforma PlaTel?

El **objeto de estudio** de la investigación de este trabajo es el reconocimiento biométrico de locutor, enfocando el **campo de acción** en los algoritmos de reconocimiento biométrico de locutor.

El **objetivo general** de esta investigación es proponer de los algoritmos de reconocimiento de locutores existentes el adecuado a la plataforma de telecomunicaciones unificada PlaTel de la Universidad de las Ciencias Informáticas.

Para lograr el cumplimiento del **objetivo** de la investigación se proponen las siguientes **tareas de la investigación:**

1. Estudiar el reconocimiento biométrico de locutores para dar una visión general sobre su funcionamiento.
2. Buscar sistemas de reconocimiento de locutores a nivel mundial para describir las tendencias actuales y los entornos de aplicación.
3. Estudiar los algoritmos de reconocimiento de locutor para profundizar en cada una de sus características.

El contenido del trabajo está estructurado en tres capítulos:

Capitulo 1 "Fundamentación teórica**"**: El contenido del capítulo hace énfasis en aspectos relacionados con la biometría, el reconocimiento biométrico de locutores y los logros que se han alcanzado tanto nacional como internacionalmente.

Capitulo 2 "Algoritmos de reconocimiento**":** En este capítulo se explican cada uno de los algoritmos de reconocimiento biométrico de locutores, enfocándose en su funcionamiento, principales ventajas y desventajas.

Capitulo 3 "Propuesta del algoritmo de reconocimiento de locutores para la Plataforma PlaTel**":** El capítulo hace referencia a algunos experimentos realizados por diferentes expertos del tema para valorar los resultados obtenidos con cada uno de los algoritmos, se tienen en cuenta diferentes aspectos relacionados con los mismos y finalmente se presenta la propuesta con una fundamentación del porqué de la selección.

CAPÍTULO 1 FUNDAMENTACIÓN TEÓRICA

Introducción

En este capítulo se abordarán aspectos relacionados con la biometría, los medios de autenticación biométrica, el reconocimiento de voz y de locutor, así como algunos logros alcanzados en diferentes áreas, tanto en Cuba, como en el mundo. Además de la evolución que han tenido los diferentes algoritmos de reconocimiento.

1.1 Biometría

El poder identificar a las personas se ha convertido en una parte muy importante en la sociedad, en aspectos como la seguridad, la banca, el control de acceso, el comercio electrónico, la prestación de servicios, el servicio social y muchos otros que requieren de la identificación de las personas. Esto puede hacerse exigiéndoles a los individuos que presenten algo que solamente ellos tienen, puede ser una llave, que es única para abrir la puerta de una casa, se pueden identificar también requiriéndoles que sepan algo, puede ser una clave de acceso, que se utiliza para abrir una caja fuerte o para acceder al correo electrónico, o puede pedírseles una combinación de ambas, que es este el caso de las tarjetas magnéticas, donde además de tener la tarjeta, la persona debe saber la clave de la misma.

Diariamente las personas se identifican, ya sea para encender el auto, marcar el ingreso a sus centros de trabajo, prender el celular, o para realizar transacciones financieras. Los múltiples usos de la identificación dejan ver cuán crucial e integral es en la vida cotidiana, no solo por la seguridad y el acceso, sino también en los sectores financieros, de salud, transporte, entretenimiento, gobierno, comunicación y control migratorio, entre otros.

Son estas razones las que dieron lugar a la práctica de la biometría, avanzada tecnología de seguridad que se basa en el reconocimiento de características físicas e intransferibles de cada individuo. La identificación biométrica es esencial ya que las características biológicas no pueden ser prestadas u olvidadas, no se pueden perder ni robar, son distintivas del individuo, además de que representan la identidad completa del mismo.

Las primeras investigaciones científicas comenzaron a realizarse en el siglo diecinueve, con el objetivo de encontrar un sistema de identificación de personas seguro que pudiera ser utilizado en fines judiciales.

Ya en el siglo veinte, muchos países en el mundo utilizaban las huellas digitales como sistema seguro de identificación.

En la actualidad la biometría está evolucionando rápidamente y tiene un amplio potencial que se usa en muchísimas aplicaciones, yendo más allá de la huella digital, a pesar de ser esta la técnica biométrica más madura y tecnológicamente probada, por su historia y resultados en el ámbito forense. Se incursiona también en otras características fisiológicas(estáticas) como son: la geometría de la mano, el reconocimiento facial, el iris y la retina, las venas de la mano, además de otras características de comportamiento(dinámicas) como son: la firma, la dinámica del teclado, el paso y la voz, convirtiéndose en un método confiable en todos aquellos lugares donde sea necesario una gran seguridad o control y que además requieran de la identificación de las personas, evitando los fraudes, el robo de información y el acceso no autorizado a redes y computadoras.

1.2 Áreas en las que se aplican sistemas biométricos.

Muchos son los países que utilizan sistemas biométricos en diferentes áreas, en la salud se emplea como mecanismo de control de acceso a historias médicas, para controlar y suministrar los medicamentos y en la prestación y autorización de servicios, por ejemplo en la Clínica Santa María de Chile se utiliza un sistema biométrico en la atención ambulatoria, evitando la suplantación de los pacientes. En Europa y Venezuela el gobierno lo utiliza en las votaciones para identificar a los votantes. En muchos almacenes son utilizados para detectar a los clientes más frecuentes y atenderlos en cuanto llegan al establecimiento. En uno de los sectores donde más se implementa la biometría es en el bancario y financiero, para evitar los fraudes y robos en las transacciones. En los aeropuertos se emplea mucho para el control fronterizo y de emigrantes.

Son muchas las ventajas de la biometría como mecanismo de seguridad, y cada día alcanza mayor grado de confiabilidad y exactitud, situándose en un puesto ventajoso frente a otros medios de seguridad como las tarjetas, que pueden ser extraviadas o las claves, que pueden ser olvidadas, sustraídas o compartidas, aunque se hace necesario mencionar que presenta algunas desventajas con respecto a estas técnicas, por ejemplo: si llegara a ser comprometida alguna clave o tarjeta, la solución es remplazarla, mientras que la biometría si llegara a comprometerse no hay posibilidad de renovarla, de igual manera, muchas personas utilizan diferentes claves para diferentes cuentas, si la clave de una cuenta falla, las demás

cuentas no se verán afectadas, pero si una biometría se quiebra entonces todas las cuentas basadas en esa biometría quebrarán.

1.3. Reconocimiento de voz.

Entre las características de comportamiento que diferencian a las personas se encuentra la voz, esta es una característica biométrica no intrusiva de gran aceptabilidad que juega un papel sumamente importante para las diversas esferas en las que se utiliza el reconocimiento biométrico del habla, otra modalidad de identificación biométrica que se ha convertido en una alternativa atrayente en los últimos años a nivel mundial, por su gran ubicuidad y su adecuada relación costo-beneficio, ya que dejando atrás las posibilidades tecnológicas, cabe destacar que es el medio de comunicación por excelencia más natural y popular.

Hace algunos años el reconocimiento de voz solo existía en la imaginación de las personas, en las películas, o en ciertos experimentos, pero en la actualidad eso ha quedado anulado por los más recientes adelantos técnicos, siendo una buena opción para la mejora de los servicios de atención a clientes, ya que hoy cualquier persona puede hacer uso de un sistema de reconocimiento muy fácilmente sin limitaciones, pues la única habilidad que se requiere para poder operarlo es uno de los conocimientos más básicos y accesibles, el habla. Se puede pensar claramente en un cliente haciendo un sin número de transacciones usando como interfaz, su voz y el teléfono, o simplemente se puede llamar a una empresa y pedir con solamente decir el nombre, navegar por internet, comprar un producto, solicitar información y algún otro servicio.

Conjuntamente con esto ha surgido la voz IP, que no es más que, un grupo de recursos que hacen posible que la señal de voz viaje a través de internet, utilizando el protocolo IP. Esta señal de voz se envía en forma digital, en paquetes, en vez de ser enviada en forma análoga, concepto que ha dado origen a los sistemas de reconocimiento del locutor, que consisten en reconocer a las personas sin necesidad de supervisión humana, solo con el análisis de su voz.

1.4 Reconocimiento de locutor.

Dichos sistemas han tomado gran auge en la actualidad debido al avance en la capacidad de procesamiento, el desarrollo de nuevos algoritmos de reconocimiento y de las nuevas tecnologías en comunicaciones, principalmente la telefonía IP.

Estos sistemas se pueden clasificar en tres tipos:

- Identificación del locutor.
- Verificación del locutor.
- Seguimiento y agrupamiento de locutores.

En la identificación del locutor, este no aporta información sobre su identidad y es el sistema el que determina quién es a partir de su voz dentro de un conjunto de posibles candidatos o, si se trata de identificación en conjunto abierto, si el locutor es conocido o no por el sistema. Por el contrario, la tarea de los sistemas de verificación de locutor es determinar si el locutor es o no quién dice ser. Por último, el seguimiento y agrupamiento consiste en etiquetar qué locutor está hablando en un segmento de voz y cuándo se producen cambios de locutores. (1)

La identificación de locutores se puede utilizar para restringir a personas no autorizadas el acceso a la información. Por otro lado, el seguimiento y agrupamiento de locutores tiene su utilidad en la transcripción de noticias o reuniones, con el fin de aislar la voz de cada uno de los locutores en una grabación. La verificación de locutores tiene también numerosas aplicaciones comerciales importantes, por ejemplo, las transacciones bancarias a través del teléfono. Además de ésta, existen muchas otras aplicaciones comerciales, todas destinadas a aumentar la seguridad en la verificación de la identidad, como podría ser la gestión de identidad en centrales de atención al cliente, haciendo posible confirmar la identidad del usuario que llama y certificar las operaciones que realice como la contratación o baja de nuevos servicios. Otra aplicación podría ser el restringir el acceso a personas no autorizados a bases de datos con información confidencial de clientes. Muy importante también son las aplicaciones en el ámbito forense, puesto que se puede emplear en juicios para comprobar si la voz empleada como prueba coincide con la del acusado. (1)

Otra clasificación de estos sistemas gira en torno a la dependencia o no del texto pronunciado, pueden ser:

- Sistemas dependientes del texto.
- Sistemas independientes del texto.

En los primeros, la locución de entrenamiento y la de verificación suelen ser el mismo texto. Fundamentalmente consiste en una palabra o frase clave (contraseña) que le permite el acceso al sistema al usuario.

En estos sistemas, la contraseña es conocida por el mismo y suele ser fija, requiriendo un nuevo entrenamiento cada vez que se desea cambiar de contraseña. Un problema de estos, es que son relativamente fáciles de atacar en caso de que el impostor grabe la palabra clave pronunciada por el usuario. Para evitar este tipo de ataques se introducen los sistemas "text-prompted" o de texto solicitado, en los que el sistema además de solicitar la contraseña al usuario, solicita repetir un código o frase elegido aleatoriamente, y que por tanto, evita la posibilidad de utilizar grabaciones. Por el contrario, en los sistemas independientes de texto la locución de entrenamiento y la de test no coinciden, siendo la locución de test desconocida por el mismo. En este caso, el sistema no utiliza ningún tipo de contraseña, sino únicamente el rasgo biométrico de la voz. (1)

Para el desarrollo de los sistemas de procesado de voz es de crucial importancia contar con buenas bases de datos orales con grabaciones de diferentes locutores, o sea, modelos que representan las características del habla de cada locutor. Estos modelos se consiguen en las sesiones de entrenamiento donde cada locutor pronunciará diversas frases.

1.5 Algoritmos de reconocimiento de locutor.

Estos sistemas funcionan bajo la implementación de algoritmos de reconocimiento que son utilizados para brindar robustez, ya que el comportamiento de los sistemas de identificación y verificación de locutor se degrada significativamente en ambientes ruidosos o en condiciones acústicas adversas. Entre los algoritmos más utilizados se encuentran las Modelos Ocultos de Markov, Redes Neuronales, Cuantización Vectorial, Modelo de Mezclas Gaussianas, Máquinas de Vectores de Soporte, entre otros, cuyo objetivo es fundamentalmente compensar los efectos del ruido.

Estos métodos han evolucionado con el tiempo, en principios se comparaba utilizando plantillas de palabras, en la década del 70, posteriormente en la década del 80 para la clasificación se comenzó a utilizar la Distorsión Dinámica en el Tiempo y la Cuantización Vectorial, ya a mediados de los 90 se realizaron enfoques estadísticos con la práctica de los Modelos Ocultos de Markov y los Modelos de Mezclas Gaussianas, alcanzando los mejores resultados hasta el momento.

1.6 Aplicaciones de los sistemas de reconocimiento de locutor.

Los sistemas de reconocimiento de locutor tienen múltiples aplicaciones, además de utilizarse en la seguridad, se puede emplear para reconocer el estado anímico de los locutores. Otra esfera donde es de suma importancia la presencia de un sistema de reconocimiento de locutor es sin duda la telefonía IP, donde no solo se convertirá al ordenador en un teléfono sino que además si cuenta con la presencia de un sistema de verificación e identificación de locutor se logrará mayor seguridad.

En el comercio electrónico las tecnologías del habla serán de gran utilidad, pudiéndose realizar los pagos mediante la voz teniendo solamente un terminal telefónico para realizar las compras y con la verificación del locutor los clientes alcanzan mayor confianza en el sistema.

Las personas discapacitadas también podrán beneficiarse de esta tecnología, con sistemas domésticos controlados por voz, para el caso de las personas que presentan una limitada movilidad, la navegación por internet para invidentes o la enseñanza a mudos utilizando el reconocimiento de voz.

1.6.1 Aplicaciones de reconocimiento de locutor en el mundo.

En la actualidad se han desarrollado disímiles aplicaciones en todo el mundo, es este el caso de BioVoicePrint, producto de SeMarket, un API de verificación que mediante reconocimiento de locutor permite garantizar un acceso seguro pudiéndose realizar transacciones bancarias vía teléfono o bien a través de un micrófono por internet, también puede utilizarse para comprar productos, todo eso con gran privacidad. Permite suprimir el uso de contraseñas en muchos casos difíciles de memorizar, compatible con cualquier reconocedor de voz, admite simultáneamente 256 llamadas entrantes y está disponible además en varios idiomas, portugués, castellano, catalán e inglés.

La empresa Agnitio presenta soluciones biométricas que son usadas en la policía de países como España, Alemania, Francia, Corea, Chile, Colombia o China. Entre estas soluciones se encuentra Batvox, herramienta que se emplea mucho en el mercado forense para la identificación de la voz.

Está pensada para ser utilizada en laboratorios de acústica o por Peritos especializados. La Policía de investigaciones de Chile es la primera institución en Latinoamérica en identificar científicamente a los delincuentes a través de la voz, tecnología que está siendo utilizada con éxito en los laboratorios de diferentes cuerpos y fuerzas de seguridad e inteligencia pública en Europa y Asia. Desde mayo de 2006,

la sección sonido audiovisual Forense del laboratorio de criminalística implementó el sistema Batvox de reconocimiento biométrico de locutor, orientado específicamente para aplicaciones forenses. (2)

Agnitio también creó los productos KOVOX, BS^3 y SAIVOX (Sistema Automático de Identificación por Voz), conocido como ASIS que significa Automatic Speaker Identification System. ASIS permite realizar búsquedas fiables entre miles de voces en solo minutos, y como resultado se obtiene una lista ordenada de los individuos que más probabilidades tienen de ser la voz anónima, casi siempre, menos de cinco y en caso contrario mostrará que no se parece a ninguna de las voces almacenadas.

La Universidad Rey Juan Carlos ha creado un sistema de reconocimiento de personas por su voz. Se trata de un primer prototipo para llegar finalmente a construir un reconocedor de altas prestaciones que permitirá verificar la identidad de diferentes individuos. (3) El proyecto se llama REBLOSI (Reconocimiento biométrico de locutor basado en la fisiología y biomecánica de las cuerdas vocales para seguridad en infraestructuras) y está financiado por la comunidad de Madrid. Participan también la Universidad Carlos III, la Universidad Politécnica de Madrid, la Universidad Politécnica de Cartagena, la Universidad Cergy-Pontoise (París) y la Universidad de California en San Diego, que se encuentra entre los diez primeros centros de investigación en reconocimiento de patrones del mundo.

El centro de tecnologías del habla en la exposición "Tecnologías de seguridad-2010" en Moscú presentó una tecnología que facilita la revelación de las "huellas" de una voz en la actividad policial y criminalística. No se trata solo de la detección en una grabación exclusiva de la voz de un sospechoso como fuente de información sobre el crimen, sino que de la detección de su mínima presencia en cualquier tipo de grabaciones. La muestra es capaz de componer un sistema informativo íntegro junto a otros datos criminalísticos. Se aplica para un análisis comparativo con el registro criminalístico de fonogramas y para verificación biométrica de la voz. Su plataforma multimedia permite agregar nueva información al sistema de datos en forma de módulos digitales de otras señales biométricas, por ejemplo, huellas digitales, retina, biometría del rostro y varias características personales de los sospechosos. (4)

El centro de investigación de tecnologías del habla (CITA) o Speech Technology Center ha creado muchísimos proyectos, está el caso de VoiceNet que busca automáticamente los locutores más parecidos y las comparaciones duran menos de medio minuto, además las muestras de voz pueden ser en cualquier idioma y por cualquier canal ya sea Teléfono, Micrófono, análogo, digital, etc. Se emplea en

organizaciones gubernamentales de investigación policial y forense. VoiceNet se encuentra instalado en las oficinas de la Procuraduría General de Justicia del Estado de Sonora, en Hermosillo, México.

Por otra parte, se encuentra IKAR LAB, el único complejo de Software y Hardware del mundo que permite ejecutar investigaciones forenses completas de grabaciones de audio y voz, de forma análoga y digital, permitiendo solucionar problemas relacionados con el análisis de la información acústica en laboratorios especializados y centros de peritajes judiciales, en los servicios de investigación de incidentes aéreos y en instituciones de enseñanza y de investigación.

1.6.2 Aplicaciones de reconocimiento de locutor en Cuba.

A raíz de todos los proyectos realizados en el mundo por diferentes países, Cuba no queda exenta de ellos, por lo que lleva a cabo una serie de investigaciones en el área del reconocimiento del locutor. El Centro de Aplicaciones de Tecnología Avanzada (CENATAV), es uno de los que ha obtenido resultados en esta área, desarrollando, implementando y probando algoritmos y métodos de extracción de rasgos acústicos y dinámicos, además de su selección y normalización, la creación de los modelos de los locutores y la posterior comparación con el habla de locutores desconocidos. Cuenta con programas que permiten probar dichos métodos, así como la extracción de rasgos acústicos del habla ante la variabilidad del canal: evolución de nuevos rasgos dinámicos SDC (shifted delta cepstra). Dicho centro ha efectuado además, diversos estudios en los que se han encontrado problemas que aún no se han podido resolver tales como:

- La extracción de rasgos acústicos con una mayor robustez ante la variabilidad del canal y del locutor, y ante el ruido.
- La selección de rasgos acústicos más representativos y menos redundantes.
- Los métodos de clasificación de locutores con un mayor poder discriminativo.
- La creación de modelos que reflejen la dinámica del habla.
- Los métodos de normalización de rasgos y de resultados de comparación más robustos y efectivos. (5)

Existen otras instituciones en el país como la Universidad Marta Abreu de Las Villas y la Ciudad Universitaria José Antonio Echeverría (CUJAE), que han desarrollado investigaciones fundamentalmente en el área de procesamiento digital de la señal de voz. En la Universidad de Oriente existe un Centro de

Estudios de Neurociencias y Procesamiento de Imágenes y Señales (CENPIS) que tiene como principal objetivo estudiar y procesar señales, para el desarrollo de aplicaciones que estén vinculadas principalmente a los problemas de salud humana, su ambiente y calidad de vida, y que los resultados de estas investigaciones conduzcan a la creación de nuevos software, dispositivos y equipos médicos, o al desarrollo permanente de los ya existentes.

Se han desarrollado numerosos trabajos, entre los que se destacan:

- Estudio de métodos para ser utilizados en la clasificación de unidades de llanto infantil, entre ellos: clasificación supervisada, redes neuronales supervisadas, el mapa auto-organizado de kohonen y algoritmos de cuantificación vectorial.
- Analizador de voz.
- Algunos rudimentos para síntesis de voz en vocabulario limitado.
- Estudio de algunos parámetros acústicos de señales de pre-vocalización, llanto y lenguaje en niños con epilepsia.
- Determinación de contorno tonal de una señal de voz.
- Bpvoz: base de procesamiento de voz.

Actualmente este grupo continúa sus investigaciones y se encuentra realizando otros proyectos como:

- Análisis del llanto infantil orientado al diagnóstico. (Proyecto con el Ministerio de Salud Pública y Proyecto de Ciencia y Conciencia de la Universidad de Oriente).
- Aplicación de redes neuronales supervisadas y no supervisadas al análisis del llanto infantil.
- Análisis de voz y llanto en enfermedades neurológicas. (6)

Conclusiones

Hoy en día es una tarea compleja realizar un sistema de reconocimiento automático de locutores que sea lo suficientemente robusto ya que se ven afectados por numerosos factores, como el ruido ambiental, el ruido en el canal de transmisión, los dispositivos utilizados para la grabación de las bases de datos de voz o sea para tomar las muestras de las voces de los locutores, por ejemplo el tipo de micrófono.

Estos no son los únicos factores que hacen que el reconocimiento no sea preciso, el factor humano también tiene un enorme peso ya que la voz de los hablantes puede variar por diferentes motivos, uno pudiera ser la velocidad de las grabaciones, ya que una misma frase dicha por un mismo locutor no es igual si lo pronuncia rápido a cuando lo hace más despacio, otra es, sin duda alguna, el estado de ánimo de las personas y algo que es imposible de descartar, es el hecho de que un locutor trate de imitar la voz de otro.

Un factor que afecta significativamente el desempeño del reconocimiento sobre redes IP, es sin duda la pérdida de paquetes, esto se produce cuando los paquetes no llegan a su destino porque son descartados, puede ser por congestión en la red, errores de transmisión, encaminamiento defectuoso o llegadas tardías y en estos casos los paquetes nunca llegarán a su destino ya que las redes IP no ofrecen por sí mismas ninguna garantía de calidad de servicios.

Por estas razones se continúan desarrollando investigaciones en este campo, con el objetivo de mejorar los problemas que afectan al mismo. Los algoritmos de reconocimiento son una buena opción para lograr robustez a la hora de identificar y verificar a un individuo, por estos motivos han sido objeto de disimiles experimentos.

CAPÍTULO 2 CARACTERIZACIÓN DE LOS ALGORITMOS DE RECONOCIMIENTO

Introducción

Para lograr un sistema capaz de realizar un buen reconocimiento de hablantes, se aplican diversos métodos y algoritmos. Todos ellos guardan sus propias características, ventajas, desventajas y cada uno tiene un propósito específico, pudiendo resolver un mismo problema a diferentes niveles. En la **figura 1** se muestra un diagrama de las etapas involucradas en el proceso de reconocimiento. Primeramente se toman las muestras de voz, luego viene la fase de extracción de vectores de características, dicha fase es anterior a la de reconocimiento y lo que hace es, extraer las llamadas características espectrales, que son las que contienen información importante de la voz de los locutores, y por último la fase de reconocimiento.

En este capítulo se explicará el funcionamiento de los algoritmos utilizados para clasificar las características de la señal de voz de cada locutor en la fase de reconocimiento, profundizando en sus características más importantes, así como sus principales ventajas y desventajas, aspectos que serán de gran utilidad a la hora de seleccionar el algoritmo más favorable, que será utilizado posteriormente por la plataforma unificada PlaTel de la Universidad de las Ciencias Informáticas.

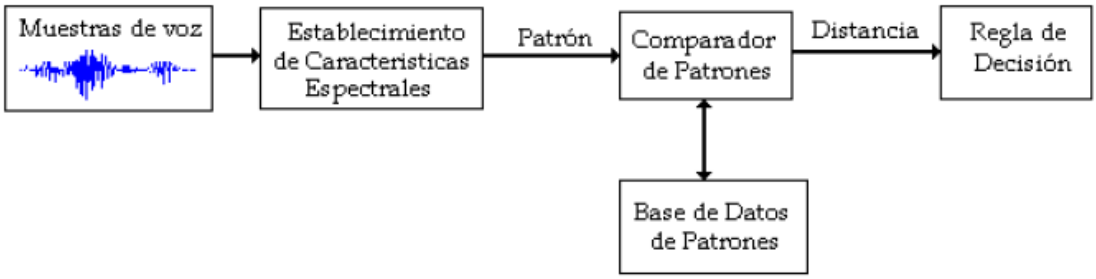

Figura 1 Etapas del reconocimiento de voz. Imagen de: (7)

2.1 Modelos ocultos de Markov

Los Modelos ocultos de Markov (HMM, por sus siglas en inglés) corresponden a modelos estocásticos que son muy usados en tecnologías dependientes del texto. Dichos modelos, al ser aplicados al reconocimiento del locutor básicamente lo que hacen es tomar la señal de voz como salida de una secuencia de estados de Markov, estos estados, serán ocultos pero observables de manera indirecta a partir de las secuencias de vectores espectrales producidos.

Un HMM puede describirse además como un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos, siendo este proceso o cadena una serie de eventos desconocidos donde la probabilidad de que ocurra uno de estos eventos depende inmediatamente del evento anterior, por lo que la cadena consta de memoria ya que los eventos futuros dependerán de los pasados, siendo el objetivo principal de estos sistemas determinar parámetros desconocidos.

Los modelos que utilizan los sistemas HMM para caracterizar la identidad de un locutor independiente del texto, son los ergódicos[1] (ver **Figura. 2**), donde no existe una ordenación correlativa[2] de las transiciones entre los distintos estados del modelo, por lo que resulta factible cualquier combinación de transición entre estados. Por otra parte, se encuentra el modelo de izquierda a derecha (ver **Figura. 3**). Este modelo solo permite transiciones hacia adelante, o sea, el índice de los estados es creciente con el tiempo de evolución, por lo que los estados del sistema van siempre de izquierda a derecha.

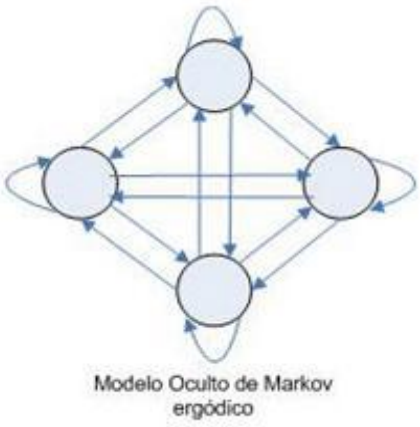

Figura 2 HMM Ergódicos.

[1] Hipótesis ergódica. Un sistema se denomina ergódico cuando la probabilidad de que el sistema se encuentre en un nodo determinado es constante para cualquier nodo escogido en un largo período de tiempo.
[2] Continua, seguida, sucesiva, etc.

Figura 3 HMM de izquierda a derecha.

Cada HMM viene dimensionado por tres ejes de referencia: la matriz de probabilidades de transición entre estados, el vector de probabilidades del estado inicial y la distribución de probabilidades de salida u observación. Respecto a esta última, existe la posibilidad de trabajar con HMM de observación discreta (DDHMM, por sus siglas en inglés) o continua (CDHMM, por sus siglas en inglés). En general, los CDHMM modelan con mayor fidelidad y producen tasas más altas de identificación que los no continuos, aunque, en condiciones de trabajo en tiempo real, los discretos parecen resultar más rápidos. (5)

Una gran ventaja de los sistemas de reconocimiento que utilizan HMM respecto a los demás es su gran versatilidad, tanto en los procesos de entrenamiento como en algunas características variables de la muestra: duración, contexto, etc. Además de mencionar su gran adaptabilidad a la variación de las condiciones del canal de transmisión.

Los HMM son ampliamente aplicados en sistemas de reconocimiento de voz, ya que con su uso, se puede realizar la agrupación en clases de las palabras o fonemas independientemente por diferente HMM, siendo discriminada cada clase fácilmente en la etapa de reconocimiento. (8)

2.1.1. Elementos que definen a los HMM.

- N: el número de estados del modelo, donde q_t denota el estado en el instante de tiempo t.
- La dimensión del conjunto de observaciones distintas de salida M, es decir, el tamaño del alfabeto $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_M}\}$
- La distribución de probabilidad de transición entre estados $\mathrm{A} = \{a_{ij}\}$:

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

- La distribución de probabilidades de emisión de símbolos entre estados $B = \{b_j(k)\}$:

  $\mathbf{b_j(O_k) = P\left(O_k | q_t = s_j\right) 1 \leq j \leq N, 1 \leq K \leq M,}$ donde O_k es un símbolo perteneciente a V.

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

  $$\mathbf{\pi_i = (q_0 = s_i) 1 \leq i \leq N}$$

Por lo que podemos definir un HMM como: $\lambda = \{A, B, \pi\}$

A raíz de esta definición surgen 3 problemas que se deben resolver para que un sistema de reconocimiento pueda utilizar los HMM en aplicaciones reales: (1)

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

Problema 1. Problema de evaluación de la probabilidad.

Dada una secuencia de observación $O = \{O_1, O_2, \ldots, O_T\}$ y un modelo $\lambda = \{A, B, \pi\}$, ¿cómo calculamos $P(O \mid \lambda)$, la probabilidad de la secuencia de observación? De ser posible calcular dicha probabilidad, se podría calcular para todos los modelos, escogiéndose aquel para el cual la probabilidad sea mayor.

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

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

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

$$P(O|Q,\lambda) = \prod_{i=1}^{T} P(O_i|q_i,\lambda)$$

Donde se asume independencia estadística de las observaciones. Por lo tanto, se obtiene:

$$P(O \mid Q,\lambda) = b_{q1}(O_1) \cdot b_{q2}(O_2) \dots b_{qT} \cdot (O_T)$$

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

$$P(Q|\lambda) = \pi q_1 \cdot a_{q1q2} \cdot a_{q2q3} \dots a_{qT-1qT}$$

que se interpreta como la probabilidad del estado inicial, multiplicada por las probabilidades de transición de un estado a otro. Sustituyendo los dos términos anteriores en el sumatorio inicial (Ecuación 1) se obtiene la probabilidad de la secuencia de observación:

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda) \cdot P(Q|\lambda) = \sum_{q1,q2,\dots qT} \pi_{q1} \cdot b_{q1}(O_1) \cdot a_{q1q2} \cdot b_{q2}(O_2) \dots a_{qT-1qT} \cdot b_{qT}(O_T)$$

A partir del resultado obtenido se interpreta lo siguiente: Inicialmente en el tiempo $t=1$nos encontramos en el estado q_1 con probabilidad π_{q1} y generamos el símbolo O_1 con probabilidad b_{q1} (O_1). Al avanzar el reloj al instante $t=2$ se produce una transición al estado q_2 con probabilidad a_{q1q2} y generamos el símbolo O_2 con probabilidad b_{q2} (O_2). Este proceso se repite hasta que se produce la última transición del estado q_{T-1} al estado q_T con probabilidad a_{qT-1qT} y generamos el símbolo O_T con probabilidad b_{qT} (O_T).

Aunque se haya logrado llegar al resultado deseado, se puede decir que no es una manera muy eficiente de calcular la probabilidad, puesto que requiere realizar $2T \cdot N^T$ operaciones, lo que por su complejidad que está en el orden de $O(N^T)$ resulta computacionalmente intratable.

Aún así existe una manera más eficiente para lograr dicho resultado. El secreto está en guardar los resultados intermedios y utilizarlos para los posteriores cálculos de la secuencia de estados. A este

algoritmo se le denomina el Algoritmo de Avance. El primer paso es definir la variable hacia delante como $\alpha t\,(i)=P\,(O_1, O_2,..., O_t, q_t = S_i \mid \lambda)$.

Esta variable corresponde con la probabilidad de que el modelo λ se encuentre en el estado i habiendo generado la secuencia parcial $O_1, O_2,..., O_t$ hasta el instante de tiempo t.

$\alpha_t\,(i)$ se puede calcular por inducción siguiendo los siguientes pasos:

1. Inicialización:

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

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

2. Inducción:

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

La expresión entre corchetes representa la probabilidad de alcanzar el estado S_j en el instante de tiempo $t+1$ partiendo de todos los estados posibles S_i en el instante t habiendo observado hasta el instante t la secuencia parcial $O_1,O_2,...,O_t$.

Si multiplicamos ahora dicho término por la probabilidad de observar O_{t+1} se obtiene $\alpha_{t+1}(j)$.

3. Finalización:

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

El cálculo de $P\ (O \mid \lambda)$ final se realiza sumando todas las variables hacia delante $\alpha_T\ (i)$ en el instante final T. Esto es así ya que por definición $\alpha_T\ (i)$ es igual a la probabilidad conjunta de haber observado la secuencia

O_1, O_2,..., O_T y encontrarnos en el estado S_i: $\alpha T\ (i) = P\ (O_1, O_2, ..., O_T, q_T = S_i \mid \lambda)$, con lo que si sumamos dicha probabilidad para todos los estados posibles obtenemos la probabilidad esperada $P\ (O \mid \lambda)$.

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

Problema 2. Problema de encontrar la secuencia de estados óptima.

La decodificación de un HMM se basa en encontrar la secuencia de estados óptima, dada una secuencia de observación. Esto resulta muy importante para tareas de segmentación y reconocimiento de voz.

Este problema tiene la particularidad que se puede resolver de diferentes maneras, mientras que en el problema 1 se podía dar una solución exacta, aquí existen múltiples formas de realizarlo. La razón es que la definición de secuencia óptima no es única, sino que existen varios criterios de optimización.

El **algoritmo de Viterbi** es el que utiliza el criterio más extendido para esto. Básicamente lo que trata es de encontrar la mejor secuencia de estados, es decir, maximizar la probabilidad $P\ (q \mid O, \lambda)$, o lo que es equivalente a maximizar $P\ (O \mid q, \lambda)$.

Dicho método se puede utilizar en la práctica para evaluar los HMM.

Para encontrar la mejor secuencia de estados $Q = \{q1, q2, ...qT\}$ para una secuencia de observación dada $O = \{O_1, O_2, ...O_T\}$ definimos la variable:

$$\delta_t(i) = \max_{q1,q2,...,qt-1} P[q_1 q_2 \dots q_t = i, O_1\ O_2 \dots O_t \mid \lambda]$$

que representa la secuencia de estados con mayor probabilidad en el instante t que acaba en el estado S_i generando las t primeras observaciones.

A continuación se muestra como el algoritmo de Viterbi selecciona y recuerda el mejor camino:

1. Inicialización

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

$$\phi_1(i) = 0$$

Inicialmente se define la probabilidad $\delta_1(i)$ como la probabilidad de encontrarse en el estado S_i en el instante t=1 multiplicada por la probabilidad de generar el símbolo O_1. El vector ϕ, en el que se va a almacenar el argumento que maximiza $\delta_t(j)$ para cada valor de t y de j, toma inicialmente el valor 0.

2. Recursión

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

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

3. Finalización

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

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

La iteración del punto 3 se termina cuando se han generado las T observaciones.

4. Backtracking

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

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

Problema 3. Entrenamiento de un modelo.

En este problema se plantea cómo deben ajustarse los parámetros del modelo $\{A, B, \pi\}$ para maximizar la probabilidad de la secuencia de observación dado el modelo $P(O \mid \lambda)$.

El mayor inconveniente que presenta este problema es que no existe un método analítico conocido que maximice el conjunto de parámetros a partir de los datos de entrenamiento. Para su solución se puede utilizar un procedimiento iterativo como el algoritmo de Baum-Welch, conocido también como (avance-retroceso), el cual utiliza los mismos principios que el algoritmo EM (Expectation Maximization) y consiste en actualizar los pesos de forma iterativa para poder explicar mejor las secuencias de entrenamiento observadas.

Es necesario definir la probabilidad hacia atrás como mismo se definió la probabilidad hacia delante antes de describir el algoritmo Baum-Welch, donde:

$\beta_t(i) = P(O_{t+1}), O_{t+2}, \dots, O_T, q_t = S_i \mid \lambda$. $\beta_t(i)$, es en este caso la probabilidad de generar la observación parcial $O = \{O_{t+1}, O_{t+2}, \dots, O_T\}$ desde el instante $t+1$ hasta el instante final T dado que el modelo se encuentra en el estado S_i en el instante de tiempo t.

$\beta_t(i)$, se puede calcular por inducción como sigue:

1. Inicialización:

   $$\beta_T(i) = 1, 1 \leq i \leq N$$

2. Recursión:

   $$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j), \quad t = T-1, T-2, \dots, 1, 1 \leq i \leq N$$

   La relación entre α y β adyacentes se puede observar mejor en la siguiente figura. α, se calcula recursivamente de izquierda a derecha mientras β se calcula recursivamente de derecha a izquierda.

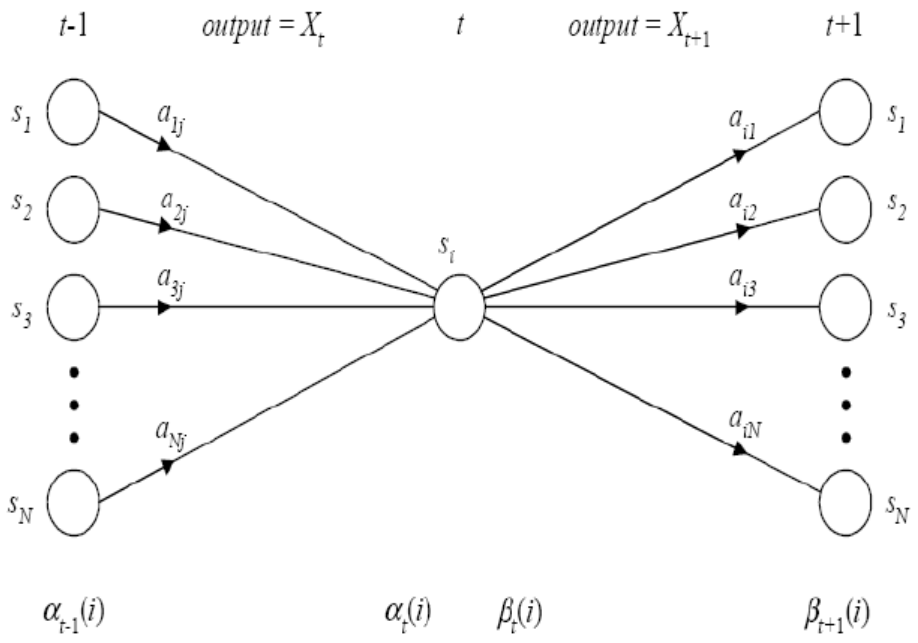

Figura 4 La relación entre α_{t-1} y α_t y β_{t-1} y β_t en el algoritmo Forward- Backward.

A continuación se define la variable $\gamma_t(i,j)$, que representa la probabilidad de realizar una transición del estado S_i al estado S_j en el instante de tiempo t dado el modelo y la secuencia de observación, es decir:

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

$$= \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i) \cdot a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)}{\sum_{k=1}^{N} \alpha_T(K)}$$

Este resultado se puede ilustrar mejor con la siguiente figura:

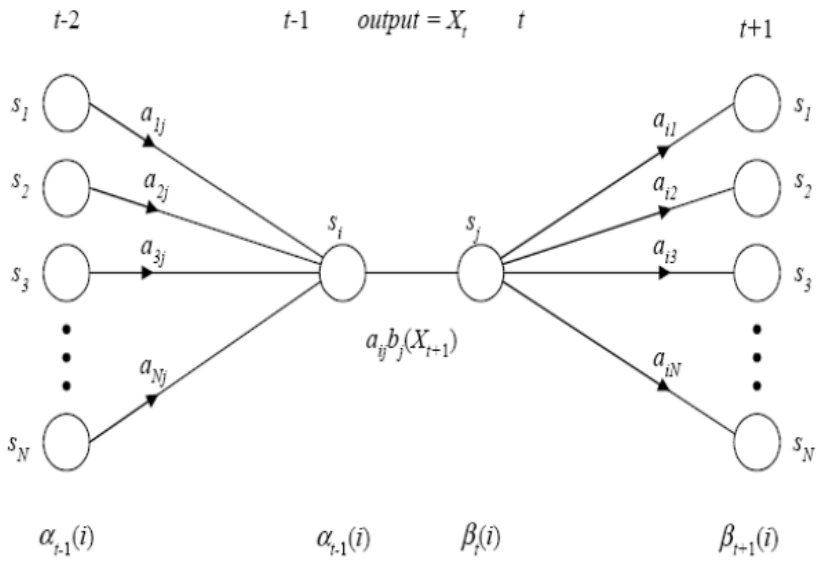

Figura 5 Ilustración de las operaciones necesarias para el cálculo de γ_t(i,j).

Es posible refinar iterativamente el vector de parámetros del HMM $\lambda = \{A, B, \pi\}$ si se maximiza la probabilidad de la observación, $P(O|\lambda)$ en cada iteración. Para ello denotamos como al nuevo vector de parámetros $\hat{\lambda}$ calculado a partir del vector de parámetros λ, obtenido en la iteración anterior. De acuerdo con el algoritmo EM, esto es equivalente a maximizar la siguiente función Q:

$$Q(\lambda, \hat{\lambda}) = \sum_{S1,S2,\ldots,SN} \frac{P(O,S|\lambda)}{P(O|\lambda)} \log P(O,S|\hat{\lambda})$$

Donde $P(O,S|\lambda)$ y $\log P(O,S|\hat{\lambda})$ se definen como sigue:

$$P(O,S|\lambda) = \prod_{t=1}^{T} a_{t-1t} b_t(O_t)$$

$$\log P(O,S|\hat{\lambda}) = \sum_{t=1}^{T} \log a_{t-1t} + \sum_{t=1}^{T} \log b_t(O_t)$$

Por lo tanto, la ecuación inicial se puede rescribir de la siguiente manera:

$Q(\lambda, \hat{\lambda}) = Q_{ai}(\lambda, \hat{a}_i) + Q_{bj}(\lambda, \hat{b}_j)$ donde

$$Q_{ai}(\lambda, \hat{a}_i) = \sum_i \sum_j \sum_t \frac{P(O, q_{t-1} = i, q_t = j \mid \lambda)}{P(O \mid \lambda)} \log \hat{a}_{ij} \quad \textbf{(1)}$$

$$Q_{bj}(\lambda, \hat{b}_j) = \sum_j \sum_k \sum_{t \in o_t = V_k} \frac{P(O, q_t = j \mid \lambda)}{P(O \mid \lambda)} \log \hat{b}_j(V_k) \quad \textbf{(2)}$$

Como se ha separado la función en tres términos independientes, se puede maximizar maximizando cada $Q(\lambda \mid \hat{\lambda})$ uno de los términos por separado, sujeto a las siguientes restricciones:

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

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

Además, los términos en las ecuaciones 1 y 2 tienen la siguiente forma:

$$F(x) = \sum_i y_i \log x_i$$

Donde:

$$\sum_i x_i = 1.$$

Haciendo uso de los multiplicadores de Lagrange, se demuestra que la función $F(x)$ toma su valor máximo en:

$$x_i = \frac{y_i}{\sum_i y_i}$$

A partir de todo esto, se obtienen las estimaciones de los parámetros del modelo HMM:

$$\hat{a}_{ij} = \frac{\frac{1}{P(O \mid \lambda)} \sum_{t=1}^{T} P(O, q_{t-1} = i, q_t = j \mid \lambda)}{\frac{1}{P(O \mid \lambda)} \sum_{t=1}^{T} P(O, q_{t-1} = i \mid \lambda)} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_t(i,k)}$$

$$\hat{b}_j(V_k) = \frac{\frac{1}{P(O \mid \lambda)} \sum_{t=1}^{T} P(O, q_t = j \mid \lambda) \cdot \delta(O_t, V_k)}{\frac{1}{P(O \mid \lambda)} \sum_{t=1}^{T} P(O, q_t = j \mid \lambda)} = \frac{\sum_{t \in O_t = V_k} \sum_{i} \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_{i} \gamma_t(i,j)}$$

Conforme con el algoritmo EM, Baum-Welch garantiza una mejora monótona en la probabilidad en cada iteración hasta que ésta converge en un máximo local.

Este algoritmo se puede resumir en los siguientes pasos:

1. **Inicialización:** Se escoge una estimación inicial del modelo λ.
2. **Paso E:** Se calcula la función auxiliar $Q(\lambda \mid \hat{\lambda})$ a partir de λ.
3. **Paso M:** Se calcula $\hat{\lambda}$ de acuerdo con las ecuaciones de re estimación para maximizar la función auxiliar Q.
4. **Iteración:** λ pasa a tomar el valor de $\hat{\lambda}$ y se repite el algoritmo desde el paso 2 hasta que converge.

2.2 Modelos de Mezclas de Gaussianas.

Los Modelos de Mezclas de Gaussianas (GMM, por sus siglas en inglés) son una técnica muy usada en sistemas independientes del texto y se basa en el modelado de los parámetros de entrada al sistema mediante modelos de mezcla de gaussianas multidimensionales, o sea, modela los diferentes vectores de parámetros dada una locución, efectuando una suma ponderada o mezcla de funciones de densidad de probabilidades gaussianas. Estos modelos, reúnen todas las virtudes de VQ, además de considerarse un HMM de un solo estado.

Los GMM no precisan en la fase de entrenamiento segmentar en estados ni entrenar la matriz de probabilidades de transiciones, aquí se obtienen los parámetros del modelo que mejor se ajusta a cada locutor, es decir, se entrena un modelo por cada locutor que constará de sus parámetros más representativos. Además, en la etapa de reconocimiento, no será necesario buscar la secuencia de estados de máxima verosimilitud, sino que bastaría con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada (5) y en la fase de test se decidirá si las locuciones de entrada se corresponden con los modelos mediante el cómputo de una medida de similitud entre ambos.

Los GMM presentan dos características que motivan a su uso, una de ellas es su capacidad de modelar algunas partes importantes del conjunto de clases acústicas de un hablante y es que el espacio acústico de un hablante se caracteriza por un conjunto de amplios eventos fonéticos como pueden ser las vocales, este conjunto de clases acústicas refleja de manera general configuraciones del tracto vocal que son completamente dependientes de cada hablante, algo que es sumamente útil para caracterizar la identidad de cada locutor. Otro aspecto poderoso que presentan los GMM es su capacidad para modelar aproximaciones precisas de densidades de formas arbitrarias, además de tener una fuerte ventaja en el manejo de las bases de datos de voz, no teniendo que hacer modificaciones en los modelos entrados en caso que aumentara el número de locutores en la base de datos, simplemente lo que hace es crear un nuevo modelo para ese locutor y después de entrado lo que resta es anexarlo al sistema completo.

2.2.1 Descripción del Modelo.

El modelo de densidades de mezclas gaussianas representa una suma pesada de M (número de mezclas) componentes de densidad descritas por: [6]

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x)$$

Donde:

- $\mathcal{X}$ es un vector D-dimensional y la matriz de rasgos

$$X = (x_1, x_2, \ldots, x_T) = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,T} \\ \vdots & \ddots & & \vdots \\ c_{D,1} & c_{D,2} & \cdots & c_{D,T} \end{pmatrix}$$

Es una sucesión de variables aleatorias indexadas por una variable discreta, el tiempo (t=1,...., T). Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y asumiremos que son independientes.

- b_i ($\mathcal{X}$) son las componentes de densidad, con i =1,2,..., M. Cada componente es una función gaussiana de la forma:

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i),\right\}$$

donde μ_i y Σ_i son el vector de medias y la matriz de covarianzas correspondientes a la i-ésima mezcla y se obtienen de la matriz de rasgos X.

- p_i son los pesos de las mezclas i =1,2,..., M y satisfacen la condición $\sum_{i=1}^{M} p_i$ =1.

- $\lambda = \{p_i, \mu_i, \Sigma_i\}$ con i =1,...., M, representa el modelo de las mezclas gaussianas donde se encuentran los parámetros que caracterizan al locutor.

Para identificar al locutor, cada uno es identificado por un modelo λ de mezclas gaussianas.

Los GMM pueden tener diferentes formas, dependiendo de la elección de la matriz de covarianza:

1. El modelo puede tener una matriz por cada componente de densidad, como el modelo descrito anteriormente (covarianza por nodo).
2. Una matriz de covarianza para todas las componentes gaussianas (covarianza grande).
3. Una sola matriz para todos los modelos que representan a los locutores (covarianza global). (5)

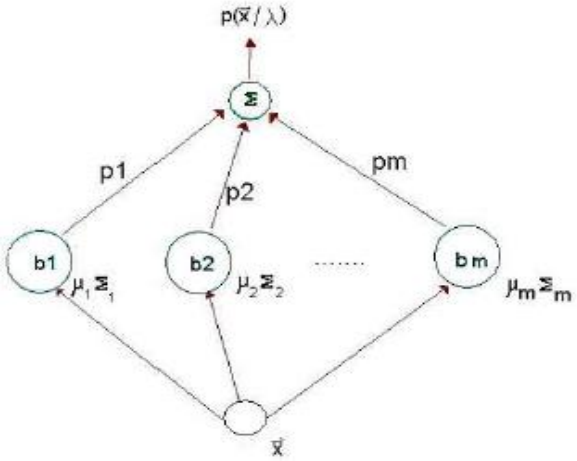

Figura 6 Modelo de mezclas gaussianas de M componentes. Imagen de: (5)

2.3 Redes Neuronales.

Las Redes Neuronales (ANN por sus siglas en inglés), son modelos computacionales que emulan la forma en que se comporta el cerebro, utilizando para ello topologías que reflejan la interconexión de las células nerviosas. Este método consiste en un grupo de neuronas que se distribuyen en capas y se interconectan por medio de caminos pesados. Cada una de estas neuronas es un elemento procesador que entrega una salida a partir de múltiples entradas, esta salida es controlada casi siempre por una función de activación. Una vez que la señal de voz es procesada digitalmente, una red neuronal artificial, permite que no haya necesidad de establecer reglas o realizar análisis estadísticos complejos para determinar el fonema en proceso de reconocimiento; o sea, debido a la capacidad de aprendizaje de las RNA, solo se necesita de una fase de entrenamiento en la cual se le presentan a la red, ejemplos característicos de determinados fonemas con sus correspondientes valores deseados, siendo el objetivo del entrenamiento brindarle a la red un conjunto de pesos que le permitan clasificar los patrones adecuadamente, dividiendo en dos etapas la operación de las redes, de forma combinada o sola, donde cabe mencionar el mecanismo de aprendizaje, que es el proceso por el cual una red modifica sus pesos dando respuesta a alguna información de entrada. En la **figura 7** se muestra un modelo de red neuronal artificial.

Una ventaja significativa que presenta este método para resolver problemas es que no es necesario tener un proceso bien definido para transformar algorítmicamente una entrada en una salida, más bien, lo que necesitan la mayoría de las redes es una colección de ejemplos representativos de la traducción que se

desea realizar. Dicho algoritmo es robusto en el sentido de que siempre se obtendrá alguna salida, incluso, en el caso de que se le presenten entradas que nunca haya visto, como tramas que contengan ruido, ya que, poseen la capacidad innata para enfrentarse a tramas ruidosas o distorsionadas con respecto a las soluciones algorítmicas tradicionales.

La capacidad de aprendizaje, estructura de cálculo distribuido-paralelo, tolerancia a fallos, adaptabilidad y plasticidad, son características que posee una RNA, a través de las cuales, puede llegar a resolver problemas que necesitarán gran cantidad de tiempo en ordenadores convencionales, además de poseer gran velocidad y robustez, ya que el conocimiento adquirido se encuentra repartido por toda la red, de forma que si se daña una parte, se continúan generando cierto número de respuestas correctas. Dentro de las principales tareas de una RNA se encuentra la de clasificar datos, es decir, debe identificar si una entrada pertenece a una clase o a otra.

Las RNA, han sido muy usadas en sistemas de reconocimiento, mostrando gran eficiencia en la clasificación de las características, además de, buen desempeño en la identificación de los hablantes, pero aún así, presentan una gran desventaja, ya que si se desea agregar uno o más hablantes al sistema de reconocimiento, una vez que se haya puesto en operación, éstas, en la mayoría de los casos, necesitan estimar todos sus parámetros nuevamente, lo que significa, que necesitan reestrenar otra vez su modelo, utilizando todas las características tanto de los hablantes ya entrenados como las de los nuevos hablantes que requieren ser agregados al sistema, razón por la cual su uso ha sido limitado. (8)

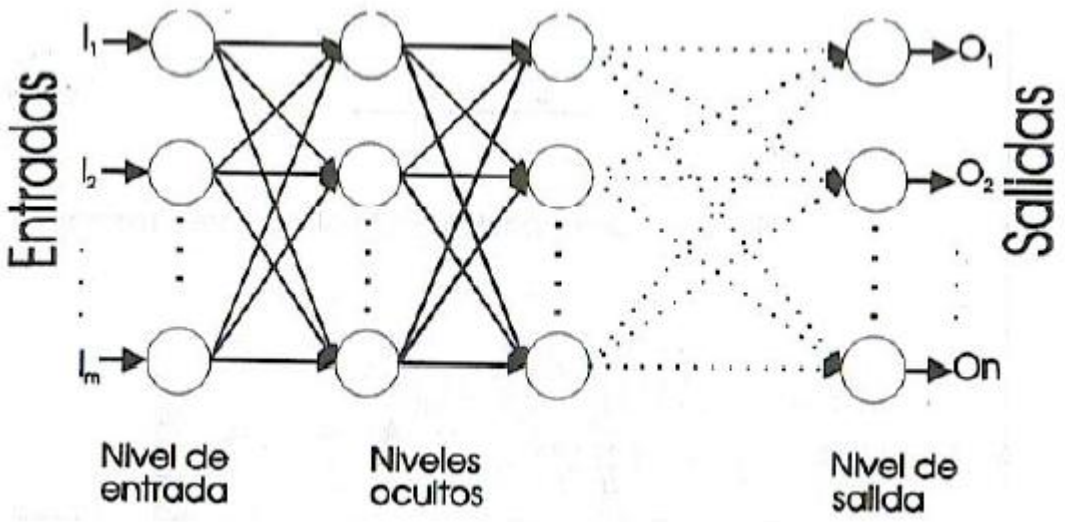

Figura 7 Modelo de una ANN. Imagen de: (9).

2.3.1 Tipos de reglas que se aplican.

Supervisado: Esta regla tiene la misión de incorporar a un maestro externo y/o información global, se divide en:

- **Aprendizaje estructural:** Lo importante aquí es encontrar la mejor relación posible entrada-salida para cada par de patrones.
- **Aprendizaje temporal:** Se captura una secuencia de patrones necesarios para alcanzar un resultado final, aquí la respuesta actual depende de las entradas y respuestas anteriores.
- **Aprendizaje por refuerzo:** Este aprendizaje se basa en la idea de no indicar en el entrenamiento la salida que se desea que la red proporcione ante una determinada entrada, en este aprendizaje el supervisor solamente indica mediante una señal de refuerzo si la salida que se obtiene se ajusta a la que se desea (deseada =+1, incorrecta =-1), en función de esto se ajustan los pesos utilizando un mecanismo de probabilidades.
- **Aprendizaje estocástico:** En este aprendizaje los valores de las conexiones de la red sufren cambios aleatorios y se evalúa el efecto partiendo del objetivo deseado y de distribuciones de probabilidad.

- **Recuperación:** Después que se establecen las conexiones, la red debe estimularse para lograr llevar a cabo el trabajo que se desea que realice, para conseguir esto la red se alimenta con un estado inicial y después de esto se procede a encontrar la información deseada. Esta etapa trae los datos asociativos que se encuentran en la memoria.

No supervisado:

Es llamado también auto-organización y no incorpora a un maestro externo, se basa solamente en información local. Existe además el aprendizaje fuera de línea donde se usan patrones para condicionar las conexiones antes de ser usada en la red, mientras que un aprendizaje en línea ocurre si algún nuevo patrón requiere ser incorporado en las conexiones de la red.

2.3.2 Ventajas de las Redes Neuronales

- **Aprendizaje adaptativo:** Las redes neuronales aprenden ciertas tareas mediante un entrenamiento, diferencian patrones y pueden modificar sus pesos, además de tener la posibilidad de cambiar constantemente para poder adaptarse a cualquier cambio.

- **Autoorganización:** La red completa puede modificarse para lograr un objetivo específico, provocando la generalización, que no es más que el derecho que tienen las redes de responder apropiadamente cuando se le presentan situaciones a las que no se habían expuesto antes.

- **Tolerancia a fallos:** Aprenden a reconocer patrones que presente ruido y pueden fácilmente seguir su funcionamiento aunque se destruya parte de la red. Su tolerancia a fallos se debe a que tienen la información distribuida entre las conexiones que existen entre las neuronas.

- **Operación en tiempo real:** Los computadores neuronales pueden ser realizados en paralelo y se diseñan y construyen máquinas con hardware para obtener esta capacidad y se puede lograr gracias a que el ajuste de los pesos de las conexiones de las neuronas es mínimo.

- **Fácil inserción dentro de la tecnología existente:** Esto posibilita la integración modular en los sistemas existentes.

El paralelismo, la velocidad y entrenabilidad que presentan, las hacen tolerantes al error, rápidas y muy eficientes a la hora de manejar grandes cantidades de datos, aunque la certidumbre, el poder de cómputo

y la lógica, no sean sus puntos fuertes, mencionando además, que cuando resuelven un problema no se logra saber como lo hacen.

2.3.3 Modelos de Redes Neuronales.

Existen diferentes modelos de redes neuronales que se utilizan en dependencia de la aplicación a desempeñar, entre los que se encuentran: Back Propagation, Perceptrón Multicapa, entre otros. A continuación se abordará con más detalles sobre la red de retro-propagación, ya que por sus características es considerada una red potente, además de ser muy usada para el reconocimiento de personas por medio de la voz, clasificando las características extraídas, para identificar o verificar a los hablantes, siendo esto, el objetivo principal de esta investigación. (7) (8)

- ✓ **Red Neuronal Back Propagation.**

La red neuronal Back Propagation es una red recurrente debido a que sus conexiones también lo son. Dentro de sus principales características se encuentra la estabilidad que presenta la misma, aunque necesitan más de un ciclo para lograrla. Su funcionamiento esencial consiste en asimilar un conjunto de datos predefinidos, los cuales están ordenados en pares de entradas-salidas, empleando para ello un ciclo de propagación del error-adaptación de pesos, o sea, el objetivo fundamental es estimular la primera capa (Capa de Entrada) con un patrón de entrada que presenta los datos que serán asimilados por la red, dicha información se propaga por la capa intermedia hasta llegar a la capa de salida, el resultado que se obtiene se compara en el que se desea obtener (Salida Deseada), calculándose un valor de error para cada neurona de salida.

Después de todo este procedimiento, estos errores son regresados de las neuronas de la capa de salida a todas las restantes de la capa intermedia, obteniendo el porcentaje de error de la participación de las neuronas intermedias en la salida original.

Dicho proceso se repite hasta que todas las neuronas reciban un error que muestre su aportación al error total. Según lo que se haya recibido, se reajustan los pesos de conexión, de forma tal que la próxima vez que se tenga el mismo patrón, la salida que se obtuvo, esté más cerca de la salida deseada, lo cual indicará que el error ha disminuido.

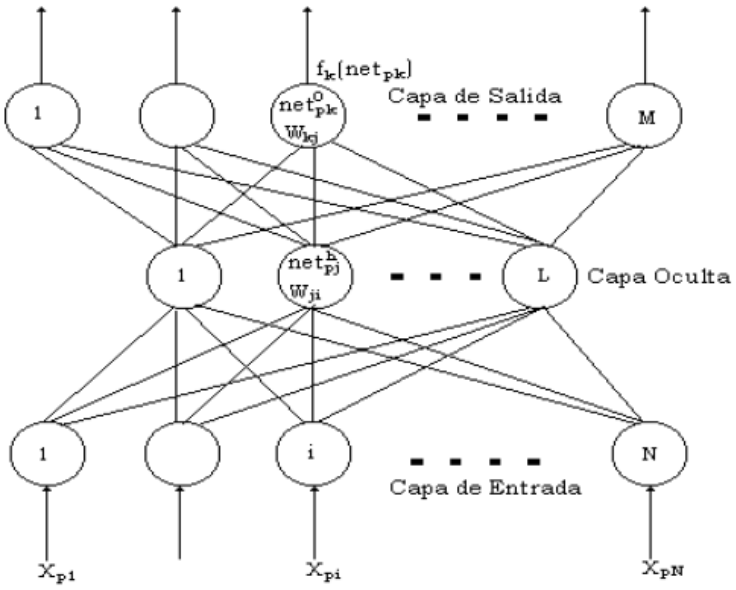

Figura 8 Arquitectura de la red neuronal Backpropagation. Imagen de: (7)

El algoritmo de entrenamiento y aprendizaje de dicha red, se resume en los siguientes pasos:

Paso 1.

Inicializar los pesos de la red con valores pequeños de aleatorios.

Paso 2.

Presentar un patrón de entrada, $X_p : x_{p1}, x_{p2}, \dots, x_{pN}$ y especificar la salida deseada que debe generar la red $d_1, d_2, \dots, d_M$ (si la red utiliza como un clasificador, todas las salidas deseadas serán cero, salvo una, que será la de la clase a la que pertenece el patrón de entrada).

Paso 3.

Calcular la salida actual de la red, para ello se presentan las entradas a la red y se calcula la salida que presenta cada capa hasta llegar a la capa de salida, ésta será la salida de la red $y_1, y_2, \dots, y_M$.

Para las neuronas de la capa de salida:

$$\gamma_{pj}^{o} = \sum_{i=1}^{N} w_{ji}^{o} x_{pi} + \theta_{j}^{o}$$

Para las neuronas de la capa oculta:

$$x_{pj}^{h} = \sum_{i=1}^{N} w_{ji}^{h} x_{pi} + \theta_{j}^{h}$$

Paso 4.

Calcular los términos de error para todas las neuronas.

$$E(n) = \frac{1}{2}\sum_{j}\left[d_j(n) - y_j(n)\right]^2$$

Paso 5.

Actualización de los pesos.

$$w_{ij}(n) = w_{ij}(n) - \alpha \Delta w_{ij}(n)$$

Donde α es el factor de convergencia y $\Delta w_{k,i}(n)$ está dado por el gradiente instantáneo:

$$\frac{\partial E(n)}{\partial \Delta w_{k,i}(n)}$$

$$\frac{\partial E(n)}{\partial \Delta w_{k,i}(n)} = \sum_{j}\left(\frac{\partial E(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial u_i(n)}\right)\frac{\partial u_i(n)}{\partial w_{k,i}(n)}$$

Y

$$u_i(n) = f_s(\sum_{k} w_{k,i} x_k)$$

$$y_j(n) = f_s(\sum_{i} v_{i,j}(n) u_i(n))$$

Paso 6.

El proceso se repite hasta que el error resulte aceptablemente pequeño para cada uno de los patrones aprendidos.

$$E_p = \frac{1}{2}\sum_{k=1}^{M} \delta_{pk}^2$$

2.4 Cuantización Vectorial

Cuantización Vectorial (VQ por sus siglas en inglés), es un método que se basa en el aprendizaje no supervisado, donde se agrupa automáticamente cada clase conocida. Dentro de sus ventajas se encuentra que permite etapas subsecuentes en el reconocedor, posibilitando así que su complejidad disminuya.

En este algoritmo se representa cada vector de entrada por el más cercano codevector o centroide de un grupo de vectores o codebook crecidamente representativos de la distribución de vectores de entrada en el espacio. Dicho codebook se selecciona con los mejores representantes de los diversos grupos en los que hayan sido divido los datos de entrada. El conjunto de centroides de cada locutor es el codebook que representa al mismo. Para la creación de los codebook se utilizan algoritmos de entrenamiento, que son supervisados o no. En el caso de los no supervisados, los codebook de cada locutor se entrenan de forma independiente, estos son los más usados porque no requieren control del entrenamiento, mientras que en los supervisados, las relaciones que presentan los codebook se tienen en cuenta para disminuir los solapamientos que pudieran presentarse.

Para aplicar VQ en el reconocimiento de locutor, es necesario crear un codebook para cada locutor de la base de datos. En el caso particular de la identificación se computa la distorsión de los vectores de rasgos de entrada con respecto a cada codebook y la menor distorsión es lo que identificará al locutor identificado. En el caso de la verificación se computa la distorsión de los vectores de rasgos de entrada pero con respecto al codebook del locutor que clama su identidad, comparando esto con un umbral, si la distorsión es menor que el umbral, es aceptado como verificado el locutor.

Los sistemas de reconocimiento de locutor que utilizan VQ, utilizan diferentes elementos que le son necesarios, estos son:

- Los datos de entrenamiento, que no son más que un conjunto de vectores de rasgos que son obtenidos de la parametrización de la base de datos de los locutores.
- El algoritmo de agrupamiento mediante el cual se dividen los datos del entrenamiento en clusters, que se representan en el codebook con sus correspondientes centroides.
- La asignación de vectores se encarga de determinar el procedimiento de búsqueda del centroide más cercano para cada uno de los vectores de rasgos de entrada de cada locutor desconocido. Generalmente se utiliza la búsqueda del vecino más cercano.
- La medida de similitud es una medida básica de distorsión que permite la asignación de los vectores de rasgos de entrada a los diferentes grupos o clusters.

2.5 Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM por sus siglas en inglés, ´´ Support Vectors Machine´´), son muy usadas en sistemas independientes del texto, que está ganando gran popularidad como herramienta para la identificación de sistemas no lineales, debido a que SVM está basado en el principio de minimización del riesgo estructural (SRM por sus siglas en inglés, ´´Structural Risk Minimization´´). Esta técnica constituye un conjunto de algoritmos de aprendizaje discriminativo, que pueden ser considerados como técnicas alternativas para el entrenamiento de clasificadores con funciones de base radial o polinomiales. Como idea principal de este algoritmo se encuentra, la de separar las clases por medio de una superficie que maximice el margen entre ellas. Este método es básicamente un algoritmo de clasificación de patrones binarios, cuyo propósito es asignar cada patrón a una clase, podemos citar el ejemplo de las ovejas, donde se tienen dos conjuntos uno de ovejas negras y otro de ovejas blancas, SVM tratará de diferenciar las ovejas por su color (clase), clasificando cada una de ellas en el conjunto blanco o en el negro.

Utiliza una función de pérdidas que intentará minimizar a la vez que se maximice el margen para tener en cuenta el solapamiento que se puede producir debido a las distorsiones en el canal de transmisión, como el ruido y otros efectos no deseados que imposibiliten un reconocimiento con robustez.

Algunas razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo solo depende de los vectores de soporte (SV por sus siglas en inglés, ´´Support Vectors´´) que son los datos con más información.

2.5.1 Las grandes ventajas de SVM son:

- Su excelente capacidad de generalización, debido a la minimización del riesgo estructural.
- El modelo solo depende de los datos que presentan mayor información por lo que presenta pocos parámetros a ajustar.
- La estimación de los parámetros se realiza optimizando una función de costo conversa por lo que se evita la existencia de un mínimo local.
- La solución que presenta SVM es sparse, lo que significa que la mayor parte de las variables son cero en la solución de SVM, esto quiere decir que el modelo final se puede escribir como una combinación de un número pequeño de vectores de entrada que son los vectores de soporte.

2.5.2 Descripción del modelo

El proceso de clasificación consiste en realizar una separación de los conjuntos de un conjunto C en diferentes subconjuntos C_i, $i = 1,\ldots, P$, denominados clases, con base en la dimensión de las características que los elementos de C poseen. Una vez que se determinan las propiedades de los subconjuntos en los que se clasificara al conjunto original (modelos), los elementos de éste son comparados con cada uno de los modelos, para establecer a cuál de ellos pertenecen. Matemáticamente este proceso puede entenderse como una función que mapea el conjunto C al conjunto de clases $\{C_i\}_{i=1}^{P}$. Se parte de la hipótesis de que, sin importar la naturaleza del conjunto C, sus elementos pueden ser clasificados de forma numérica.

Esta representación puede ser en $\mathbb{R}^n$, para algún n € $\mathbb{N}$. Sin embargo, bajo este planteamiento, la labor de clasificación no guarda dificultad alguna, ésta surge, por ejemplo, cuando se considera que los elementos del conjunto C son resultado de un conjunto finito de variables aleatorias $\mathbb{R}^n$ denotado por:

$$C = \{X_1, X_2, \ldots, X_k\},$$

donde X_j es una variable aleatoria discreta infinita o continua. Si consideramos que las clases C_i son una partición de C entonces, dadas las hipótesis, el proceso de clasificación puede no ser exhaustivo, por lo que de manera práctica, C es formado con subconjuntos de valores representativos de cada una de las variables aleatorias X_j. De aquí que se pueda decir que el objetivo de una Máquina de Soporte Vectorial, consiste en modelar en cierta forma el comportamiento de cada una de las variables aleatorias X_j, de forma que se pueda determinar, dado un vector propuesto, a cuál de ellas pertenece.

En particular, para la clasificación de voz puede considerarse, sin pérdida de generalidad, que el conjunto C está formado por dos variables aleatorias, es decir, que

$$C = \{X_1, X_2\}.$$

Es posible representar a cada elemento del conjunto C, de la siguiente forma:

$$(x_j, y_j), j = 1, \dots, \iota$$

Donde $x_i \epsilon \mathbb{R}^n, y_j \epsilon \{-1, 1\}$ y ι es la cardinalidad de C. Suponiendo que se toma una muestra representativa de cada una de las variables aleatorias, la representación dada previamente permite establecer al conjunto C de la siguiente forma:

$$C = C1 \cup C2, \qquad C1 \cap C2 = \emptyset,$$

$$C1 = \{x_1, \dots, x_k\},$$

$$C2 = \{x_{k+1}, \dots, x_\iota\}.$$

A raíz de las consideraciones que se han hecho, la distribución de los puntos que conforman a C_1 y C_2 es desconocida a priori. Por lo que se pueden considerar dos casos, cuando C_1 y C_2 son linealmente separables y cuando no lo son.

1. **Conjuntos Separable Linealmente.**

Se puede decir que C_1 y C_2 son linealmente separables cuando existe un hiperplano en $\mathbb{R}^n$ determinado por un vector w perpendicular al mismo de forma que

$$\varpi \cdot x + b = 0,\ b \in \mathbb{R} \qquad (3)$$

Para cualquier punto x en el hiperplano y además

$$\varpi \cdot x_i + b > 0,\ \forall x_i \in C_1, \qquad (4)$$

$$\varpi \cdot x_j + b < 0,\ \forall x_j \in C_2. \qquad (5)$$

Se puede decir que si C_1 y C_2 son linealmente separables entonces la existencia de un hiperplano tal, determinado por un vector ϖ, no es única, ya que existe una infinidad de esos vectores. Por lo que es necesario establecer un criterio que permita determinar cuál de ellos se tomará para la clasificación. Además, si ambos conjuntos son separables entonces existe una distancia mínima entre los mismos. La siguiente figura presenta un esquema del caso que se muestra.

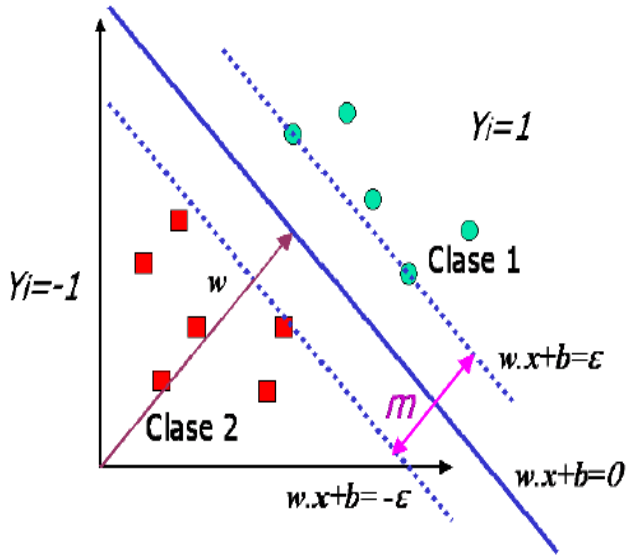

Figura 9 Esquema de dos conjuntos de vectores linealmente separados mediante un hiperplano que maximiza el margen m. Imagen de: (10)

De la esta, se puede observar la construcción de dos hiperplanos paralelos al original, determinado éste por ϖ, los cuales delimitan un margen entre los conjuntos y cuya magnitud m podemos relacionar con

dichos hiperplanos de la siguiente forma: Supongamos que ϖ está contenido en el hiperplano inferior entonces sucede que

$$\varpi \cdot \varpi + b = -\varepsilon \qquad (6)$$

Para alguna constante $\varepsilon > 0$. Entonces el vector dado por:

$$\varpi + \frac{\varpi}{\|\varpi\|} m \qquad (7)$$

Está incluido en el hiperplano superior, lo cual implica que:

$$\varpi \cdot \left(\varpi + \frac{\varpi}{\|\varpi\|} m\right) + b = \varepsilon. \qquad (8)$$

Restando las ecuaciones 6 y 8 y simplificando obtenemos:

$$\|\varpi\| m = 2\varepsilon \Rightarrow m = \frac{2\varepsilon}{\|\varpi\|}. \qquad (9)$$

Dada la igualdad (9) se infiere que si deseamos maximizar la magnitud del margen m es necesario minimizar la magnitud de ϖ. Retomando nuevamente la notación ($\mathcal{X}_j$, $\mathcal{Y}_j$) con $\mathcal{Y}_j \in \{1, -1\}$, para cada uno de los vectores $\mathcal{X}_j$ de C, entonces el problema de encontrar un hiperplano que separe a C_1 y a C_2, maximizando el margen de separación entre dichos conjuntos, queda planteado como:

$$\min\left\{f(\varpi) = \frac{1}{2\varepsilon}\|\varpi\|^2\right\} \qquad (10)$$

$$y_i(\varpi \cdot x_i + b) \geq \varepsilon,$$

$$i = 1, \ldots, \iota,$$

$$\varepsilon > 0, b \in \mathbb{R}.$$

Este es un problema de optimización de tipo cuadrático sujeto a ι restricciones en $\mathbb{R}^n$ y cuya solución puede darse a partir del uso de la teoría de multiplicadores de Lagrange en n variables.

Denotemos las restricciones que aparecen en el problema (10) de la siguiente forma:

$$g_i(\mathfrak{w}) = y_i(\mathfrak{w} \cdot x_i + b),\ i = 1, \dots, \iota, \qquad (11)$$

Si denotamos por $\mathfrak{w}_k$ a la k- ésima entrada de $\mathfrak{w}$ sabemos que se cumple

$$\frac{\partial f}{\partial \mathfrak{w}_k}(\mathfrak{w}) = \sum_{i=1}^{\iota} \alpha_i \frac{\partial g_i}{\partial \mathfrak{w}_k}(\mathfrak{w}),\ k = 1, \dots, \iota \qquad (12)$$

$$\alpha_i g_i(\mathfrak{w}) = \varepsilon,\ i = 1, \dots, \iota, \qquad (13)$$

Para ciertas constantes $\alpha_i \in \mathbb{R}$ a determinar. Sustituyendo en la ecuación (12) a la función f y a las funciones g_i se tiene el siguiente desarrollo:

$$\frac{\partial f}{\partial \mathfrak{w}_k} = \frac{1}{2} w_k = \sum_{i=1}^{\iota} \alpha_i \frac{\partial g_i}{\partial w_k} = \sum_{i=1}^{\iota} \alpha_i y_i x_i(k) \Rightarrow \qquad (14)$$

$$\mathfrak{w} = \varepsilon \sum_{i=1}^{\iota} \alpha_i y_i x_i.$$

Por otro lado, de la igualdad (13) se tiene que

$$\frac{\partial \sum_{i=1}^{\iota} \alpha_i g_i(\mathfrak{w})}{\partial b} = 0 \Rightarrow \sum_{i=1}^{\iota} \alpha_i y_i = 0.$$

Por lo tanto, la solución al sistema (10) está dada por

$$\mathfrak{w} = \varepsilon \sum_{i=1}^{\iota} \alpha_i y_i x_i \qquad (15)$$

$$\sum_{i=1}^{\iota} \alpha_i y_i = 0 \qquad (16)$$

Para obtener el valor de cada una de las constantes α_i sustituimos la ecuación (15) en las restricciones iniciales (16) con lo que se obtiene la siguiente ecuación:

$$\alpha_i y_i \left(\varepsilon x_j \cdot \sum_{i=1}^{\iota} \alpha_i y_i x_i + b \right) = \varepsilon, \qquad j = 1, \dots, \iota. \qquad (17)$$

Derivando parcialmente cada una de las ecuaciones dadas en 17 con respecto a α_j, se obtiene que

$$\alpha_j \|x_j\|^2 + y_j \sum_{i=1}^{\iota} \alpha_i y_i x_i \cdot x_j = -\frac{y_j b}{\varepsilon}, \qquad j = 1, \dots, \iota. \qquad (18)$$

Las igualdades descritas en **ecuación (18)** conforman un sistema lineal de ι ecuaciones con incógnitas $\alpha_i, i = 1, \dots, \iota$ por lo que se tiene un sistema de la forma $A\alpha = B$, con $A \in \mathbb{R}^{lxl}$ y $B \in \mathbb{R}^l$, siendo α el vector de incógnitas. Este sistema tiene solución única sólo si A tiene inversa. Es importante recordar que el valor ι es la cardinalidad del conjunto C, por lo que la complejidad en la obtención de una solución mediante alguna implementación, dependerá del orden de dicho conjunto. Los valores de las constantes α_i quedarán en términos de b y $\mathcal{E}$, de los cuales uno de ellos puede ser propuesto y el otro determinado, por ejemplo, con la **ecuación (6).**

Así, el vector ϖ dado en la **ecuación (15)**, es conocido como el vector de soporte del hiperplano que separa a C_1 y a C_2, de donde deriva el nombre de Máquinas de Soporte Vectorial.

2. Conjuntos No Separable Linealmente.

Se puede considerar que el sistema lineal de ecuaciones dado en **ecuación (18)**, al tratar el caso de dos conjuntos C_1 y C_2 separables linealmente es una condición. Aún así, dicho sistema también puede ser obtenido para dos conjuntos cualesquiera de $\mathbb{R}^n$. Por lo que se puede decir que C_1 y C_2 no son separables linealmente cuando la matriz correspondiente al **sistema (18)** no tenga solución, en ese caso no es posible la construcción de un hiperplano que cumpla las condiciones del problema planteado en **ecuación (10).**

El procedimiento para el caso de dos conjuntos no separables linealmente consiste en utilizar una función $\phi: \mathbb{R}^n \to \mathbb{R}^m$ con $n, m \in \mathbb{N}, m \geq n$ ó $m = \infty$. Esta función mapea a los conjuntos $C_1 = \{x_1, \dots, x_k\}$ y $C_2 = \{x_{k+1}, \dots, x_j\}$, a un espacio de mayor dimensión, donde se pueden denotar por $\Gamma_1 = \{\phi(x_1), \dots, \phi(x_k)\}$ y

$\Gamma_2 = \{\phi(x_{1k+1}), \ldots, \phi(x_i)\}$ respectivamente. Dicho mapeo se realiza con el objetivo de que los conjuntos obtenidos Γ_1 y Γ_2 sean separables linealmente o en su caso se minimice el error mediante la separación con un hiperplano, o sea, que el número de vectores clasificados incorrectamente sea mínimo. **La figura 10** esquematiza la operación ideal de la función ϕ sobre un conjunto dado.

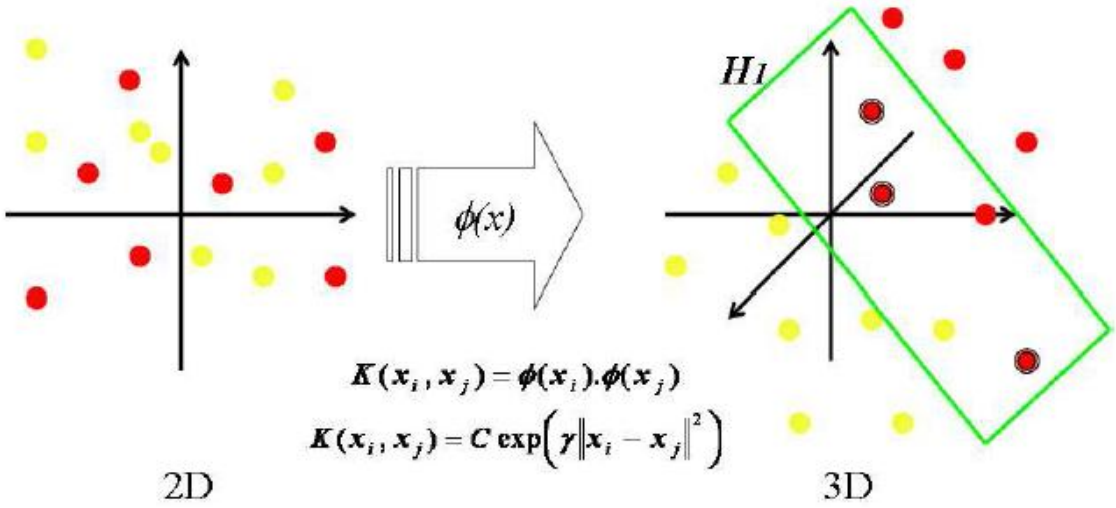

Figura 10. Esquema de la transformación de dos conjuntos no separables linealmente mediante la función. Imagen de: (10)

La función ϕ mapea los conjuntos de vectores no separables linealmente a un espacio de mayor dimensión. Estos vectores son separables linealmente y la solución al problema original mediante el procedimiento explicado en la sección **2.5.2** parte **1**, es un hiperplano $H_1 \subset \mathbb{R}^m$. En este caso se considera que existen vectores de los conjuntos Γ_1 y Γ_2 que se encuentran contenidos en H_1. A dichos vectores se les conoce como vectores de soporte. Para obtener una solución en el espacio $\mathbb{R}^n$ original, se realiza el mapeo inverso de los vectores de soporte, los cuales determinarán las fronteras que separarán a los conjuntos C_1 y C_2. Se considera que los vectores que determinan estas fronteras conforman el modelo para C_1 (o equivalentemente para C_2).

No se puede determinar a priori, dados dos conjuntos de vectores $C_1, C_2 \subset \mathbb{R}^n$, una función ϕ que cumpla los objetivos descritos previamente, por lo que el procedimiento para determinarla no es constructivo. Por

tanto, el tratamiento para este caso está basado en la realización de ensayos con funciones ϕ conocidas. Así, la función ϕ es de especial importancia en la solución del problema de clasificación.

Del procedimiento dado en **2.5.2** parte **1**, puede observarse que las operaciones con vectores involucran el producto punto o producto interno canónico en $\mathbb{R}^n$. Este proporciona una función que determina una norma y a su vez una métrica para el espacio:

$$\|x\| = \sqrt{x \cdot x}, \qquad (19)$$

$$d(x, y) = \|x - y\|. \qquad (20)$$

Tales normas y métricas, respectivamente, son empleadas también al separar los conjuntos Γ_1 y Γ_2 en $\mathbb{R}^m$. Así que el problema análogo al **(10)**, planteado en este nuevo espacio es:

$$\min\left\{f(\mathfrak{w}) = \frac{1}{2}\|\mathfrak{w}\|^2, \quad \mathfrak{w} \in \mathbb{R}^m\right\} \qquad (21)$$

$$y_i(\mathfrak{w} \cdot \phi(x_i) + b) \geq \varepsilon,$$

$$\varepsilon > 0, b \in \mathbb{R}.$$

Por lo anterior la ecuación análoga a **(18)** en este nuevo espacio está dada por

$$\alpha_j\|\phi(x_j)\|^2 + y_j\sum_{i=1}^{\iota}\alpha_i y_i \phi(x_i)\cdot\phi(x_j) = -\frac{y_j b}{\varepsilon}, \quad j = 1, \dots, \iota. \qquad (22)$$

Se infiere de la **ecuación (22)**, que la función ϕ puede ser vista como una que modifica la norma y métrica del espacio original dadas en **(19 y 20)**, por las siguientes:

$$\|x\|_\phi = \sqrt{\phi(x)\cdot\phi(x)}, \qquad (23)$$

$$d_\phi(x, y) = \|x - y\|_\phi \qquad (24)$$

A la función definida por

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \qquad (25)$$

Se le conoce como función núcleo y su uso es más importante que el de la propia función ϕ, de la cual no se requiere su conocimiento en forma explícita, ya que es suficiente, como lo muestra la **ecuación (22)**, con establecer la función núcleo K para obtener una solución.

Algunos ejemplos de funciones núcleo que han sido sugeridas o empleadas en problemas de clasificación son las siguientes:

Lineal:

$$K(x_i, x_j) = x_i \cdot x_j.$$

Polinomial:

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \quad \gamma > 0,\ r, d \in \mathbb{R}.$$

Función de Base Radial:

$$K(x_i, x_j) = c\ exp(-\gamma \parallel x_i - x_j \parallel^2), \quad \gamma > 0,\ c \in \mathbb{R}.$$

Sigmoide:

$$K(x_i, x_j) = tanh(\gamma x_i \cdot x_j + r),\ \gamma, r \in \mathbb{R}.$$

Aunque existen diferentes núcleos es común el uso de la Función de Base Radial, por los resultados obtenidos durante la clasificación. Sin embargo, puede optarse por el uso de otros núcleos dependiendo de los resultados obtenidos para un caso particular. (10)

2.6 Distorsión Dinámica en el Tiempo.

La Distorsión Dinámica en el Tiempo (DTW, por sus siglas en inglés), es un método basado en ajustes de plantillas que ha sido muy usado en sistemas de reconocimiento de locutores dependientes del texto, comparando diferentes realizaciones temporales de las mismas expresiones, o sea, compara la locución

de entrada con un conjunto de plantillas que no son más que un conjunto de rasgos acústicos ordenados en el tiempo y que representan las expresiones a reconocer. Dichas expresiones son almacenadas en las plantillas durante la fase de entrenamiento. En el reconocimiento se alinea de manera óptima la secuencia de datos de entrada con el modelo de referencia almacenado con anterioridad. Una vez concluida la comparación, la distancia acumulada entre las dos expresiones es la base de la puntuación. Si las expresiones son idénticas en el tiempo, el trayecto de alineamiento es una diagonal; si no son idénticas, las desviaciones de la diagonal representan las distancias requeridas a distorsionar. En esta técnica la mayor puntuación de semejanza o menor distancia se obtiene para diferentes realizaciones de un mismo *password* de un mismo locutor.

Este método no requiere muchos recursos computacionales en la fase de entrenamiento y es bastante simple, precisamente por esta simplicidad es fácilmente aplicable en tareas como control de acceso con *password*, teniendo previamente las plantillas de todos los posibles *passwords* para cada locutor autorizado. Esto es una gran desventaja que presenta este algoritmo, es altamente dependiente de las expresiones de referencia, imposibilitando la variabilidad en la señal de voz. Esto hizo que estos sistemas fueran poco flexibles, llevando a DTW a caer en desuso.

Dentro de sus ventajas se encuentra la gran eficiencia que brinda el mismo, dando la oportunidad de comparar dos señales que en tiempo son distintas pero en contexto son iguales, ofreciendo mayor exactitud en el reconocimiento. A continuación se muestra el algoritmo para calcular la distancia:

Lo más importante es alinear de forma óptima la secuencia de vectores de parámetros de entrada $T = \{t_1, t_2,\ldots, t_N\}$ con el modelo de referencia $R = \{r_1, r_2,\ldots, r_M\}$, donde N es en general distinto de M debido a la variabilidad de la duración ya comentada antes. Necesitándose una función que relacione las N muestras de la secuencia de entrada y las M de la plantilla, minimizando la distorsión entre ambas. La función será de la forma $m = W(n)$ y debe de cumplir además las siguientes restricciones:

- $W(1) = 1$
- $W(N) = M$.

Dadas dos secuencias cualesquiera, la función $W(n)$ es el camino de alineamiento óptimo entre ambas y se obtiene resolviendo la siguiente ecuación:

$$D^* = min\left\{\sum_{n=1}^{N} d\left[t_n, rw_{(n)}\right]\right\}$$

donde $d\ [t_n, rw_{(n)}]$, es la distancia entre el instante n de la secuencia de entrada y el instante W(n) de la plantilla. Al final del alineamiento, D^* es la distancia acumulada sobre el camino óptimo W(n) entre R y T constituyendo la base para la puntuación resultante, en la que además se pueden incluir costes adicionales que penalicen caminos que sean demasiado no diagonales. (1)

2.7 Resumen de las ventajas y desventajas más importantes.

	Ventajas	Desventajas
Distorsión Dinámica en el Tiempo (DTW)	Detectan y comparan tramos fonéticos de alta estabilidad (vocales abiertas, consonantes nasales) aplicando técnicas de correlación cruzada, coherencia, entre otras, para la medida de distancias. Estos sistemas DTW han sido utilizados en algunas metodologías forenses como un complemento a otros análisis clásicos.	Los principales inconvenientes de estos sistemas se relacionan con la enajenación de la información a nivel suprasegmental y la necesidad de supervisión en las tareas de segmentación.
Cuantización Vectorial(VQ)	Reducción sensible de la capacidad de almacenamiento en el cálculo del análisis espectral y una reducción de la complejidad computacional en el cálculo de distancias (se	Sus inconvenientes más significativos están relacionados con la distorsión espectral por el error de cuantificación (al representar cada vector por un

	puede usar cálculos tan simples como la distancia euclideana o la de Mahalanobis).	representante).
Red Neuronal artificial(ANN)	Las redes son robustas al ruido y permite tener en cuenta el contexto de la señal, pueden crearse redes que tengan un funcionamiento similar a las VQ, a las GMM, HMM y otros algoritmos en el reconocimiento del locutor.	Presentan como limitantes que la mayoría de las redes requieren almacenar todos los datos del entrenamiento durante la clasificación, requiriendo, en algunos casos, un volumen apreciable de memoria y poder de cálculo.
Modelos Ocultos de Markov(HMM)	Su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. A todo ello, hemos de añadir su gran adaptabilidad a la variación de las condiciones de voz o del canal de transmisión y lógicamente, su funcionalidad en condiciones dependientes de texto.	Alto costo computacional, y sus mejores resultados se encuentran en el reconocimiento del locutor dependiente del texto.
Modelo de mezclas	Las GMM pueden representar, con un alto grado de fidelidad,	Presentan un alto costo computacional implicando un

Gaussianas(GMM)	un amplio margen de distribuciones muestrales, como es el caso de los diferentes coeficientes cepstrales que puede generar una locución. Además de las ventajas citadas, interesantes estudios comparativos sobre el rendimiento de diferentes técnicas de reconocimiento automático, ante distintas circunstancias (procesos de entrenamiento, factores de degradación, etc.), han contribuido a tomar como mejor sistema básico a las GMM.	tiempo considerable al crear los modelos.

Tabla 1 Resumen de las ventajas y desventajas más importantes. Tabla de: (5)

CAPÍTULO 3: PROPUESTA DEL ALGORITMO DE RECONOCIMIENTO DE LOCUTOR PARA LA PLATAFORMA PLATEL.

Introducción

En este capítulo se realizará un análisis de algunos experimentos y pruebas realizadas por expertos en el tema para ver los resultados obtenidos con el uso de los diferentes algoritmos de reconocimiento del locutor. Se evaluará la evolución que han tenido estos métodos desde sus inicios hasta la actualidad, profundizando en aquellos algoritmos que tienen su mejor funcionamiento en sistemas de reconocimiento dependientes del texto. Finalmente, se expondrá el algoritmo que en correspondencia con todo lo planteado anteriormente y en el transcurso de esta investigación, es el más adecuado para ser utilizado en el reconocimiento de locutor que se desea incorporar en la plataforma PlaTel.

Para seleccionar la propuesta del algoritmo de reconocimiento, se tuvieron en cuenta 3 aspectos fundamentales:

1. Los algoritmos que han alcanzado mejores resultados en sistemas de reconocimiento automático de locutores dependientes del texto, por aportar un reconocimiento más seguro que los sistemas que funcionan bajo texto independiente, buscando con esto que los servicios brindados por PlaTel alcancen la seguridad requerida.
2. Los que fueran más fuertes ante vulnerabilidades que pueden presentarse en estos sistemas a la hora de la identificación y verificación, dígase pérdida de los paquetes, degradación de la señal de voz por los efectos del ruido o las condiciones del canal de transmisión y demás factores que imposibilitan un buen reconocimiento.
3. Los más utilizados en los últimos años.

3.1 Experimentos y pruebas realizados por expertos del tema.

- Resultados obtenidos por (7) con la red neuronal y los modelos de mezclas gaussianas, donde se obtuvo un mayor desempeño para la primera, con un porcentaje de reconocimiento de 99.12%, mientras que en GMM se obtuvo un 84.25%. Los resultados se muestran en la **figura 11** y **12** respectivamente.

	Patrones Prueba	Hablante 1	Patrones Prueba	Hablante 2	Patrones Prueba	Hablante 3	Patrones Prueba	Hablante 4
Palabra 1	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 2	118	98.33%	115	95.83%	119	99.17%	118	98.33%
Palabra 3	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 4	119	99.17%	114	95.00%	118	98.33%	119	99.17%
Palabra 5	120	100.00%	118	98.33%	120	100.00%	119	99.17%
Palabra 6	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 7	119	99.17%	116	96.67%	117	97.50%	119	99.17%
Palabra 8	120	100.00%	120	100.00%	120	100.00%	119	99.17%
Palabra 9	120	100.00%	117	97.50%	120	100.00%	120	100.00%
Palabra 10	118	98.33%	118	98.33%	120	100.00%	118	98.33%
Porcentaje Individual		**99.55%**		**98.16%**		**99.50%**		**99.33%**
			Porcentaje Global de Reconocimiento					**99.12%**

Figura 11 Resultado de la red neuronal.

	Hablante 1	Hablante 2	Hablante 3	Hablante 4	Funcionamiento Individual
MODELO λ_1	**86.6%**	0.0%	0.0%	13.3%	**86.67%**
MODELO λ_2	0.0%	**90.0%**	0.0%	10.0%	**90.0%**
MODELO λ_3	0.0%	20.7%	**70.3%**	9.0%	**70.36%**
MODELO λ_4	0.0%	10.0%	0.0%	**90.0%**	**90.0%**
Porcentaje Global de Reconocimiento					**84.25%**

Figura 12 Resultado con el Modelo de mezclas Gaussianas.

- Experimento realizado por (10), utilizando Máquinas de Soporte Vectorial, se puede apreciar que los mejores resultados de clasificación obtenidos equivalen a 82.142%, este resultado fue comparado con el resultado obtenido por otros sistemas basados en SVM y las cifras son muy similares.
- Experimento publicado en (11), utilizan RNN, HMM y DTW donde el resultado individual de cada algoritmo es del 80%, mientras que al unirlos el promedio fue de 97%, esta unión le da una ventaja

al reconocedor, si DTW falla, se compensa esta falla con los demás algoritmos, algo que no se puede lograr si se aplica solamente DTW. Los datos se aprecian en la **figura 13**.

Palabra Pronunciada	ANN	HMM	DTW	Modelo de Mayoría
1	80%	75%	95%	100%
2	90%	90%	85%	100%
3	80%	90%	100%	90%
4	85%	90%	95%	100%
5	95%	90%	100%	100%
6	65%	75%	100%	100%
7	85%	85%	95%	95%
8	70%	70%	95%	95%
9	85%	85%	100%	95%
0	65%	75%	80%	100%
Fin	95%	100%	100%	95%
Porcentaje de aciertos	81%	84%	95%	97%

Figura 13 Porciento de aciertos por palabras y algoritmos.

- Experimento realizado en (7), utilizando los Modelos de Mezclas Gaussianas, las pruebas se realizaron con 4, 5, 8, 10, 16 mezclas, pero el desempeño aceptable se logró con 5 mezclas, ya que con los demás valores el resultado no variaba y los tiempos de cálculos se volvían más tardíos, entrando en ocasiones en ciclos infinitos. Los datos se muestran en la **figura 14.**

	Hablante 1	Hablante 2	Hablante 3	Hablante 4	Funcionamiento Individual
MODELO λ_1	**86.6%**	0.0%	0.0%	13.3%	**86.67%**
MODELO λ_2	0.0%	**90.0%**	0.0%	10.0%	**90.0%**
MODELO λ_3	0.0%	20.7%	**70.3%**	9.0%	**70.36%**
MODELO λ_4	0.0%	10.0%	0.0%	**90.0%**	**90.0%**
Porcentaje Global de Reconocimiento					**84.25%**

Figura 14 Porcentaje de aciertos para cuatro hablantes.

- Experimento realizado en (7), para una Red Neuronal Backpropagation en un sistema dependiente del texto, donde se utilizaron 10 etapas de entrenamiento correspondientes a las 10 palabras de cada hablante que se tienen en la base datos, el desempeño más eficiente se logró utilizando 15 neuronas. Los porcentajes de acierto se muestran en la **figura15.**

	Patrones Prueba	Hablante 1	Patrones Prueba	Hablante 2	Patrones Prueba	Hablante 3	Patrones Prueba	Hablante 4
Palabra 1	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 2	118	98.33%	115	95.83%	119	99.17%	118	98.33%
Palabra 3	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 4	119	99.17%	114	95.00%	118	98.33%	119	99.17%
Palabra 5	120	100.00%	118	98.33%	120	100.00%	119	99.17%
Palabra 6	120	100.00%	120	100.00%	120	100.00%	120	100.00%
Palabra 7	119	99.17%	116	96.67%	117	97.50%	119	99.17%
Palabra 8	120	100.00%	120	100.00%	120	100.00%	119	99.17%
Palabra 9	120	100.00%	117	97.50%	120	100.00%	120	100.00%
Palabra 10	118	98.33%	118	98.33%	120	100.00%	118	98.33%
Porcentaje Individual		**99.55%**		**98.16%**		**99.50%**		**99.33%**
			Porcentaje Global de Reconocimiento					**99.12%**

Figura 15 Funcionamiento de la Red Neuronal.

- Prueba realizada por (12), donde se utiliza el método DTW, alcanzando una taza de reconocimiento del 85%, la matriz de confusión del sistema se muestra en la **Figura 16.**

n° / Pal	1	2	3	4	5	6	7	8	9	10
Uno	19	1	–	–	–	–	–	–	–	–
Dos	–	15	–	–	–	2	–	–	–	3
Tres	–	2	18	–	–	–	–	–	–	–
Cuatro	–	–	–	19	–	–	–	1	–	–
Cinco	–	–	–	–	18	–	–	–	–	–
Seis	2	–	–	–	–	18	–	–	–	–
Siete	–	–	–	–	–	3	17	–	–	–
Ocho	–	–	–	–	–	–	–	17	–	–
Nueve	1	–	–	–	–	–	–	–	19	–
Diez	–	5	3	–	–	2	–	–	–	10

Figura 16 Matriz de confusión del sistema de reconocimiento

- Prueba realizada (13) haciendo uso de un HMM, utilizando frases con confusión lingüística, y frases que carecen de esa confusión, se realizaron 20 repeticiones con 5 personas, 2 mujeres y 3 hombres, de las cuales se usaron 50% para el entrenamiento y 50% para las pruebas. El resultado se muestra en la **figura 17.**

Modelos	de Harkov
3 estados	5 estados
89.5%	95.5%
95%	97.5%

Figura 17 Porciento de reconocimiento para HMM de tres y de cinco estados, usando habla discontinua.

3.2 Modelos Ocultos de Markov.

En el transcurso de la investigación se han analizado detalladamente los algoritmos de reconocimiento, sus características, así como, sus ventajas y desventajas, rendimiento, funcionamiento. En consideración a todo lo antes planteado y a los parámetros que se tuvieron en cuenta en este apartado para la selección, se propone para la plataforma PlaTel los Modelos Ocultos de Markov, por ser el método que más se destaca dentro de aquellos que se emplean con mejores resultados en sistemas dependientes del texto, ver tabla 2. En este grupo los HMM cumplen con los parámetros planteados:

1. Lo que para algunos es una gran desventaja que presentan los HMM, para el presente trabajo de investigación es una poderosa ventaja y es precisamente que los mejores resultados que se han obtenido con este método son en sistemas de reconocimiento dependientes del texto. Esta técnica presenta una gran funcionalidad en estas condiciones, bien se observa en los resultados de los experimentos analizados, donde la mayoría de las pruebas realizadas a los sistemas que utilizan HMM demuestran un porciento de acierto en el reconocimiento superior al 95% y eso es precisamente lo que se pretende lograr en el reconocedor de PlaTel.
2. Crear un sistema automático que identifique a las personas por su voz es algo bastante complejo por todos los problemas que pueden presentarse, HMM a diferencia de otros métodos presenta gran adaptabilidad a la variación de las condiciones de la voz y del canal de transmisión y aunque no es lo suficientemente robusto ante el ruido, en este aspecto se han alcanzado muchísimos logros, existiendo técnicas que son aplicadas antes de aplicar los algoritmos para compensar los efectos del ruido, además de otras variantes que pueden emplearse, como realizar las grabaciones en ambientes ruidosos y realizar posteriormente las pruebas en un ambiente similar, por lo que el ruido no es algo que pueda afectar el buen funcionamiento de los HMM. Entre las técnicas usadas para suplir los efectos del ruido, hay muchas que lo hacen sobre los parámetros de la señal de voz, es este el caso de la substracción espectral, filtrado de Wiener, serie de Taylor vectorial (VTS), normalización en media, varianza de los coeficientes cepstrales, entre otras, diseñadas para limpiar la señal del ruido aditivo y la distorsión del canal, todo esto previo al reconocimiento. Existen además otros métodos que realizan la compensación adaptando los modelos al entorno ruidoso, con un reentrenamiento de los modelos o la combinación de modelos en paralelo.

   La pérdida de paquetes es otro de los problemas asociados a las tecnologías de voz en entornos IP y los HMM se adaptan bien al carácter rafagueante de las pérdidas, ya que tienen la capacidad de capturar la dependencia temporal entre ellas, pudiendo tener un modelo que incluya un estado para la recepción correcta de los paquetes y otro para las pérdidas. Es este el caso del Modelo Gilbert, que no es más que un Modelo de Markov que sencillamente calcula la probabilidad de que un paquete se reciba cuando el anterior se recibió correctamente y de que un paquete se reciba cuando el anterior se ha perdido. Esta alternativa de Markov consigue modelar de forma bastante fiable, el comportamiento de una red real. Un sistema de reconocimiento del locutor basado en

HMM se evita la necesidad de utilizar otro algoritmo para tratar las pérdidas, pudiendo tener uno capaz de realizar las dos funcionalidades. Esto demuestra que esta técnica está condicionada para enfrentar todas las imprecisiones que pueden presentarse en el reconocimiento de locutores.

Dependientes del texto	**Independientes del texto**
HMM	GMM
ANN	SVM
DTW	VQ

Tabla 2 Muestra en que sistemas han tenido mejor desempeño los algoritmos de reconocimiento.

3. Los algoritmos de reconocimiento han ido evolucionando en el tiempo**,** ver **figura 18,** observándose que los más utilizados en los últimos años han sido los HMM y GMM, por los considerables resultados alcanzados en sistemas de reconocimiento dependientes e independientes del texto respectivamente. Se han dado inmensos pasos en el área del reconocimiento de hablantes, primeramente se comparaba utilizando plantillas con pequeñas bases de datos, luego de introducirse grandes bases datos de habla natural, se necesitaron nuevas soluciones, es por eso que estos enfoques estadísticos, dejaron prácticamente en desuso a métodos como DTW, por diversas limitaciones que presenta en su funcionamiento. En este caso GMM no es lo que se busca, a pesar de sus buenos resultados y es porque precisamente estos son alcanzados en condiciones independientes del texto.

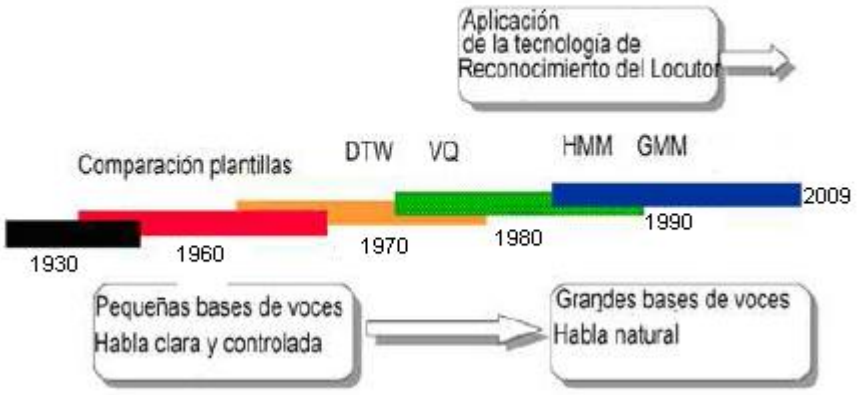

Figura 18 Evolución de la utilización de los algoritmos de reconocimiento de patrones hasta el 2009. Imagen de: (5)

Conclusiones

Los modelos Ocultos de Markov han sido ampliamente probados, y cada día surgen nuevas modificaciones a este método, todo en función de un mejor funcionamiento y la mejora de sus buenos resultados. Aún se sigue trabajando en nuevas técnicas, otra que puede ser de total consideración es el sistema Híbrido HMM-ANN, ya que intenta unir las fortalezas de los HMM con las ANN, y a su vez compensar entre ellos sus debilidades. Con los Modelos Ocultos de Markov se puede lograr un buen funcionamiento del reconocedor de PlaTel teniendo en cuenta que un buen algoritmo de reconocimiento no es lo único que se necesita para ello, pero si es uno de los aspectos más importantes.

.

CONCLUSIONES

El estudio teórico realizado permitió conocer aspectos importantes de los sistemas de reconocimiento de locutor, así como sus aplicaciones a nivel nacional y mundial, las tendencias actuales y las características de los mismos. Además de saber cuáles son los algoritmos que utilizan estos sistemas para el reconocimiento y el porqué de su selección.

Finalmente, se obtuvo un resultado después de haber realizado un estudio profundo de los diferentes algoritmos que se emplean para el reconocimiento del locutor, permitiendo así que la plataforma PlaTel pueda realizar una identificación lo más coherente y segura.

Con la culminación de este trabajo se dio cumplimiento al objetivo principal, ya que se propuso una técnica que brinda robustez a la hora de identificar y verificar al locutor, lo cual es muy importante para dicha plataforma.

RECOMENDACIONES

Al concluir el presente trabajo se recomienda:

- ✓ Continuar con el estudio de estos temas con el fin de adquirir un amplio conocimiento sobre los sistemas de reconocimiento del locutor.
- ✓ Seleccionar los algoritmos de extracción de características espectrales y compensación del ruido para lograr un sistema de reconocimiento con un mayor acabado.
- ✓ Realizar pruebas en la plataforma Platel con el objetivo de evaluar el funcionamiento del algoritmo seleccionado.

Trabajos citados

1. **Elizalde, Cristina Esteve.** *RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO MEDIANTE ADAPTACIÓN DE MODELOS OCULTOS DE MARKOV FONÉTICOS.* 2007.

2. Cienciaforense. [En línea] [Citado el: 9 de febrero de 2010.] www.cienciaforense.cl/csi/content/view/35/2/.

3. Universia. [En línea] [Citado el: 9 de febrero de 2010.] www.universia.es/portada/actualidad/noticia_actualidad.jsp?noticia=93991.

4. Biometria. [En línea] [Citado el: 9 de febrero de 2010.] www.biometría.gov.ar/noticias/nueva-herramienta-en-biometría-de-voz.aspx.

5. CENATAV. [En línea] [Citado el: 10 de febrero de 2010.] www.cenatav.co.cu/es/temas.htm(8).

6. Cenpis. [En línea] [Citado el: 12 de febrero de 2010.] www.cenpis.uo.edu.cu/gpv.htm#trab_desarr.

7. **López, Marisol Hernandez.** *Sistema para reconocimiento de hablantes dependiente e independiente del texto.* México : s.n., 2004.

8. **Acevedo, Eric Simancas.** *Reconocimiento de hablantes basado en Modelo de Mezclas Gaussianas.*

9. **Lazarini, Arturo Mora.** *Funciones Gabor Bidimensionales para el análisis y clasificación de texturas.*

10. **Bernal, Juan Gabriel Pedroza.** *APLICACIÓN DE LAS MÁQUINAS DE SOPORTE VECTORIAL AL RECONOCIMIENTO DE HABLANTES.*

11. *Reconocimiento de voz con redes neuronales, DTW y Modelos Ocultos de Markov.* **Carlos Alejandro de Luna Ortega, Julio Cesar Martinez, Miguel Mora Gonzalez.** 032, Mexico : s.n., 2006.

12. *IMPLEMENTACIÓN DE UN RECONOCEDOR DE PALABRAS AISLADAS DEPENDIENTE DEL LOCUTOR.* **César San Martín S., Roberto Carrillo A.** Chile : s.n.

13. **Rodriguez, Jose Luis Orepesa.** *Algoritmos y métodos para el reconocimiento de voz en español mediante sílabas.*

BIBLIOGRAFÍA

Elizalde, Cristina Esteve**.** *RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO MEDIANTE ADAPTACIÓN DE MODELOS OCULTOS DE MARKOV FONÉTICOS.* 2007.

Cienciaforense. [Online] [Cited: febrero 9, 2010.] www.cienciaforense.cl/csi/content/view/35/2/.

Universia. [Online] [Cited: febrero 9, 2010.] www.universia.es/portada/actualidad/noticia_actualidad.jsp?noticia=93991.

Biometría. [Online] [Cited: febrero 9, 2010.] www.biometría.gov.ar/noticias/nueva-herramienta-en-biometría-de-voz.aspx.

CENATAV. [Online] [Cited: febrero 10, 2010.] www.cenatav.co.cu/es/temas.htm(8).

Cenpis. [Online] [Cited: febrero 12, 2010.] www.cenpis.uo.edu.cu/gpv.htm#trab_desarr.

López, Marisol Hernandez. *Sistema para reconocimiento de hablantes dependiente e independiente del texto.* México : s.n., 2004.

Acevedo, Eric Simancas. *Reconocimiento de hablantes basado en Modelo de Mezclas Gaussianas.*

Lazarini, Arturo Mora. *Funciones Gabor Bidimensionales para el análisis y clasificación de texturas.*

Bernal, Juan Gabriel Pedroza. *APLICACIÓN DE LAS MÁQUINAS DE SOPORTE VECTORIAL AL RECONOCIMIENTO DE HABLANTES.*

Reconocimiento de voz con redes neuronales, DTW y Modelos Ocultos de Markov. **Carlos Alejandro de Luna Ortega, Julio Cesar Martinez, Miguel Mora Gonzalez.** 032, Mexico : s.n., 2006.

IMPLEMENTACIÓN DE UN RECONOCEDOR DE PALABRAS AISLADAS DEPENDIENTE DEL LOCUTOR. **César San Martín S., Roberto Carrillo A.** Chile : s.n.

Rodriguez, Jose Luis Orepesa. *Algoritmos y métodos para el reconocimiento de voz en español mediante sílabas.*

Medida de la calidad de voz en redes IP. ***José Joskowicz, Rafael Sotelo.***

Comunicaciones Unificadas con Elasatix. ***Edgar Landívar,*** *2008-2009.*

Reconocimiento de locutor independiente del texto (en ambientes ruidosos). ***Francisco García López, Marcos Faúndez Zanuy.***

Aplicación de RNA y HMM a la verificación automática de locutor. ***F.L Alegre.***

Identificación Biométrica de locutores para el ámbito forense. Felipe ***Ochoa, César San Martín, Roberto Carrillo.***

Selección y pesado de parámetros acústicos mediante algoritmos genéticos para el reconocimiento del locutor. ***Maider Zamalloa, Germán Bordel, Luis Javier Rodríguez, Mikel Peñagarikano, Juan Pedro Uribe.***

Efectos de la extensión del ancho de banda en reconocimiento del locutor. ***Marcos Faúndez-Zanuy.***

API (Application Programming Interface): Interfaz de Programación de Aplicaciones.

ANN (Artificial Neural Networks): Redes Neuronales Artificiales.

Biometría: Parte de la biología que estudia cuantitativamente la variabilidad individual de los seres vivos, utilizando métodos estadísticos.

BioVoiceprint (Biometric Voiceprint Authenticator): es una API de verificación de usuario que, mediante el reconocimiento de locutor, permite garantizar un acceso seguro a un sistema determinado.

Dependiente de texto: Un sistema de verificación de locutores en el que el texto que el locutor debe decir es conocido por el sistema.

DTW (Dynamic Time Warping): Distorsión Dinámica en el Tiempo.

GMM (Gaussian Mixture Model): Modelo de Mezclas Gaussianas.

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

Hylafax: Software que permite enviar y recibir faxes.

Independiente de texto

Un sistema de verificación de locutores en el que el sistema no conoce el texto que el locutor ha dicho.

Identificación: Tarea en la cual el sistema biométrico busca en una base de datos una referencia que coincida con la muestra biométrica suministrada y, de encontrarla, devuelve la identidad correspondiente. Se recopila información biométrica y se la compara con todas las referencias en la base de datos.

SVM (Support Vectors Machine): Máquina de Vectores de Soporte.

SMS (Short Message Service): Servicio de Mensajes Cortos.

Openfire: Sistema de mensajería instantánea que permite tener tu propio servidor de mensajería y administrar a tus usuarios, compartir archivos, auditar mensajes, mensajes offline, mensajes broadcast, grupos, etc.

Postfix: Servidor de correo.

SeMarket: 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, firma electrónica y protección de activos digitales, que tiene como objetivo asegurar la identidad de las personas y organizaciones, para ello cuenta con una amplia gama de soluciones que utilizan tecnologías biométricas tales como el reconocimiento de cara, voz y huella dactilar, así como servicios de firma electrónica basados en tecnología de Infraestructura de Clave Pública(PKI).

Umbral: Valor predeterminado de un usuario para las tareas de verificación o identificación de grupo abierto en los sistemas biométricos. La aceptación o el rechazo de los datos biométricos dependen de si el resultado de coincidencia se encuentra por encima o por debajo de la escala. La escala es ajustable de modo que el sistema biométrico puede ser más o menos estricto según los requisitos de cada aplicación biométrica.

Verificación: Tarea durante la cual el sistema biométrico intenta confirmar la identidad declarada de un individuo, al comparar la muestra suministrada con una o más plantillas registradas con anterioridad.

VQ (Vector Quantization): Cuantización Vectorial.

VoIP: Voz sobre IP.