RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos de transmisión de voz sobre internet: VoIP. El reconocimiento del locutor en Internet

José Ramón Calvo de Lara

RT_078 **noviembre 2015**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de transmisión de voz sobre internet: VoIP. El reconocimiento del locutor en Internet

José Ramón Calvo de Lara

RT_078 **noviembre 2015**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Métodos de transmisión de voz sobre internet: VoIP. El reconocimiento del locutor en Internet

José Ramón Calvo de Lara

Equipo de Investigaciones de Imágenes y Señales, Centro de Aplicaciones de Tecnología de Avanzada (CENATAV), La Habana, Cuba
jcalvo@cenatav.co.cu

RT_078, Serie Azul, CENATAV
Aceptado: 23 de noviembre de 2015

Resumen. El presente reporte pretende brindar información sobre las características técnicas de la trasmisión de la voz sobre el protocolo TCP/IP de Internet, analizar el comportamiento de la calidad del servicio que brinda dicho protocolo y su repercusión en el procesamiento automático del habla, específicamente en el reconocimiento del locutor. No se pretende agotar el tema, solo introducir al lector en esta tecnología, dar a conocer algunas experiencias reportadas e identificar los principales retos a que se enfrentan los métodos de reconocimiento de locutores aplicados sobre VoIP.

Palabras clave: reconocimiento de locutores, VoIP, compresión de voz, codificación de voz, voz por internet, biometría en internet.

Abstract. This report aims to provide information on the technical characteristics of voice transmission over TCP / IP Internet protocol, analyzing the behavior of the quality of service provided by the protocol and its impact on automatic speech processing, specifically in speaker recognition. It is not intended to exhaust the subject, only to introduce the reader to this technology, to present some reported experiences and to identify the main challenges that speaker recognition methods are applied to VoIP.

Keywords: speaker recognition, VoIP, voice compression, voice encoding, voice over internet, biometrics on internet.

1 Introducción

En el mundo actual, la trasmisión de voz sobre el protocolo TCP/IP[1] de la red Internet, conocida como VoIP[2], se ha convertido en soporte vital para las comunicaciones habladas. Desde las comunicaciones

[1] Transmission control protocol/internet protocol: protocolo de control de transmisión de datos por internet

[2] Voice over Internet protocol: protocolo de trasmisión de voz por internet

gubernamentales y militares hasta las personales, requieren ya de Internet para mantenerse y desarrollarse. Cuba no es una excepción, los enlaces entre plantas telefónicas están siendo transferidos a dicho protocolo y muchos de los enlaces de telefonía internacional ya se realizan sobre el mismo. Muchas empresas, instituciones y universidades cuentan con pizarras telefónicas soportadas sobre dicho protocolo, aprovechando la infraestructura de red existente en dichos lugares. La estrategia de informatización de la sociedad cubana esta soportada entre otros elementos, por una eficiente y segura utilización de la VoIP para las comunicaciones por voz, tanto alambrada como inalámbrica.

Los métodos de procesamiento automático del habla, como el reconocimiento de palabras, de locutores, de idiomas, de emociones, etc., deben enfrentarse a las nuevas condiciones que presenta la voz, debido a su codificación, compresión y trasmisión por paquetes, procesos propios del protocolo TCP/IP.

En la referencia [1] el lector puede conocer la evolución detallada de la VoIP desde sus inicio en la década del 70 aplicada sobre ARPANET[3] hasta la actualidad aplicada sobre redes WiMAX[4].

2 Las limitaciones de las redes públicas conmutadas y la alternativa de voz sobre Internet

Por muchas décadas del pasado siglo, la tecnología de conmutación de circuitos analógicos de banda estrecha fue la base de los servicios de telecomunicaciones. Pero las comunicaciones analógicas no son ni robustas ni eficientes ante el ruido y las distorsiones propias de las líneas telefónicas conmutadas y la limitación de banda de las mismas entre 300 y 3400 Hz provoca otros efectos no deseados en la voz.

Al aparecer las comunicaciones digitales, las redes públicas conmutadas PSTN[5] comienzan a procesar y trasmitir la voz muestreada a 8 kHz sobre la misma red conmutada de circuitos analógicos de banda estrecha con una velocidad de hasta 64 kb/s, utilizando el método de compresión por codificación PCM[6], que brinda la mejor calidad posible dentro de la banda telefónica de 300 hasta 3400 Hz. La voz codificada digitalmente es más robusta y flexible aunque requirió un ancho de banda extra para su trasmisión [1], por tal razón han sido utilizadas durante años otras técnicas de compresión como log-PCM[7] y ADPCM[8].

La conexión entre los puntos terminales se establece conmutando los circuitos, al comenzar la comunicación de voz y se mantiene mientras se lleva a cabo la comunicación, impidiendo que a esos circuitos puedan acceder otros servicios. Aunque se han logrado grandes avances en el aprovechamiento de dichos canales, esta arquitectura, ajustada para trasmitir voz, no es flexible si se desea lograr la convergencia de datos, video y tráfico de voz, de conjunto sobre el mismo circuito.

La aparición de Internet, brindando servicios basados en el protocolo TCP/IP, ha sido una solución al problema de la convergencia, al basarse en la conmutación y trasmisión de paquetes de datos y no de circuitos. Hoy día, la trasmisión de voz sobre dicho protocolo conocido como VoIP, constituye uno de

---

[3] Advanced Research Projects Agency Network: es la red de computadoras creada por encargo del Departamento de Defensa (DOD) de Estados Unidos en 1969, la génesis de Internet.

[4] Worldwide Interoperability for Microwave Access: es una norma de transmisión de datos por radio para la última milla o bucle de abonado, en las frecuencias de 2,3 a 3,5 GHz con una cobertura de hasta 50 km. Brinda servicios de banda ancha en zonas donde el despliegue de cable o fibra presenta costos muy elevados por la baja densidad de población.

[5] Plain Switched Telephone Networks: redes telefónicas conmutadas.

[6] Pulse Code Modulation: modulación por impulsos codificados.

[7] Modulación por impulsos codificados en escala logarítmica.

[8] Adaptive delta PCM: Modulación por impulsos codificados delta adaptado

los servicios de telecomunicaciones más prominentes y de más rápido crecimiento en el mundo, que soporta nuevas aplicaciones como centros de llamadas sobre Internet, operaciones bancarias y comerciales por vía telefónica, comunicaciones personales a larga distancia y de muy bajo costo.

Para la trasmisión de voz sobre Internet, en el extremo trasmisor la voz debe ser comprimida con algún método de codificación, las muestras codificadas son insertadas en paquetes secuenciales que son transportados a través de la red siguiendo diferentes conexiones definidas por los interfaces de conmutación y enrutamiento hasta llegar al extremo receptor donde, se decodifican y se almacenan en un buffer, sintetizándose de nuevo la voz, la que se reproduce. Un esquema muy simplificado de este proceso se muestra en la figura 1, donde se incluye en el extremo receptor la aplicación de reconocimiento de locutores ("ASR system").

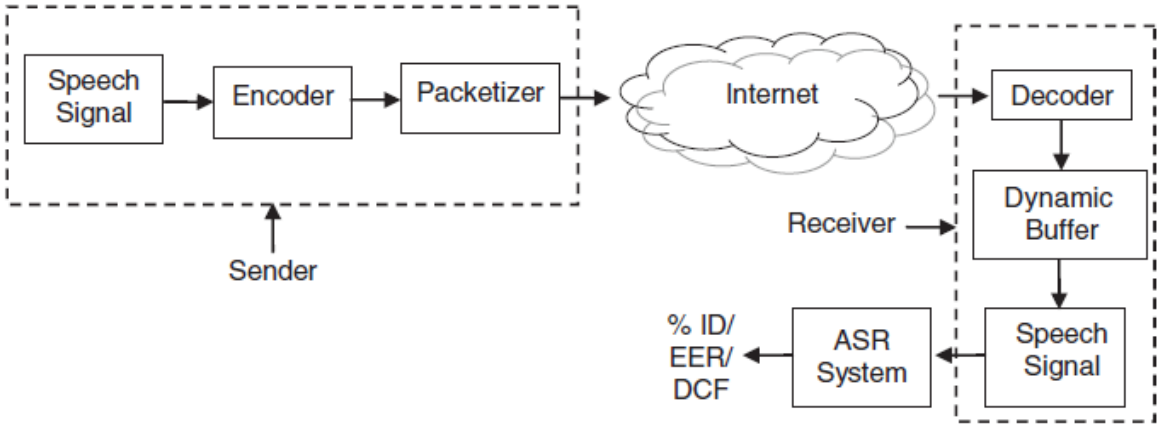


Fig. 1. Esquema del proceso de la VoIP, incluyendo el reconocimiento de locutores (ASR) [2].

La VoIP tiene varios rasgos avanzados que la caracterizan y que la hacen la mejor alternativa a las redes conmutadas PSTN, lo que ha provocado una rápida migración de las comunicaciones telefónicas hacia la telefonía por VoIP [1]:

- Bajo costo: es la principal ventaja de la VoIP, las llamadas de larga distancia pueden hacerse a muy bajo costo accediendo a las infraestructuras y conexiones ya existentes en Internet. La VoIP trata la voz como cualquier otro dato, permitiendo la mensajería de voz y la realización de video-conferencias.
- Servicios integrados: La VoIP permite su integración a la telefonía PSTN, facilitando las conexiones de las llamadas entre abonados.
- Actualizaciones fácilmente escalables: en los sistemas PSTN, un incremento en la conectividad conlleva a un incremento en el costo. En VoIP no necesariamente se requiere un costo extra para incrementar la conectividad, que puede lograrse con un incremento de las funcionalidades soportadas sobre software en una red de computadoras.
- Recuperación ante desastres: en VoIP, una llamada involucra gran número de dispositivos interconectados y ante un fallo se encuentran vías alternativas para mantener la conectividad.
- Aplicaciones avanzadas: La integración de la VoIP y la PSTN no se limita solo a llamadas telefónicas. Otros servicios relacionados con la voz como identificador de llamadas, re-direccionamiento de llamadas o envío simultáneo de mensajes, son implementados de forma simple

y pueden actualizarse cuando sea necesario. La ITU[9] además provee recomendaciones para la implementación de fax sobre VoIP.

- Seguridad: una red privada virtual consiste en la utilización de un determinado ancho de banda de Internet, donde el acceso público se restringe con la utilización del encriptado de los datos que viajan por ella, incluso la voz.

Las comunicaciones por voz soportadas sobre VoIP pueden agruparse, según el interfaz utilizado para el acceso a Internet, en:

- Telefonía PSTN integrada a TCP/IP: es la más antigua y extendida, el lazo de abonado PSTN se conecta en la planta telefónica a la plataforma Internet.
- Telefonía TCP/IP: (conocido como "IP-phone") el enlace entre abonados está totalmente soportado en Internet, muy común en las denominadas pizarras IP de instituciones, universidades y empresas.
- VoIP entre computadoras: mensajería de voz, videoconferencias, chat de voz en servicios como Skype, GoogleTalk, etc.
- Acceso inalámbrico a Internet: redes WLAN[10] y telefonía móvil GSM[11] y 3GPP[12]

3 Parámetros de calidad del servicio del protocolo TCP/IP aplicados en VoIP

La medida QoS[13] se define como la habilidad de la red para proveer un servicio que satisfaga al usuario, o sea, mide el grado de satisfacción del usuario. La QoS se ha convertido en un elemento muy sensible al trasmitir paquetes de voz, porque las aplicaciones en tiempo real como la VoIP, son muy sensibles a las demoras.

Para lograr que la VoIP sea la mejor alternativa a las redes PSTN, deben proveer a los usuarios alta QoS, similar o mejor que la de éstas. Por tal razón las aplicaciones VoIP exigen altas restricciones a la QoS de las redes IP, que en ocasiones no se alcanzan. Los principales factores que restringen la QoS del protocolo TCP/IP y que influyen decisivamente en la calidad de las comunicaciones de VoIP son [1], [2], [3]:

- Demora promedio punto- punto: conocida también como demora labios-oídos o latencia, es el intervalo de tiempo que transcurre desde el momento que el hablante expresa una palabra y el que escucha la oye.
- Demora por "jitter"[14]: es la variación inestable en la demora punto-punto entre paquetes consecutivos.
- Razón de entrega de paquetes: es la velocidad en que transcurre la comunicación, que se ve afectada debido a la pérdida de paquetes.
- Compresión de la voz: es el método de codificación - decodificación "códec" utilizado para reducir la razón de trasmisión y aprovechar mejor el ancho de banda.

---

[9] Internacional Telecomunications Union: Unión Internacional de Telecomunicaciones

[10] Wireless Local Area Network: red de área local inalámbricas. (LAN: red de área local)

[11] Global System for Mobile Communications: sistema global de comunicaciones móviles, estándar de comunicaciones móviles propuesto por el ETSI: EuropeanTelecommunications Standards Institute

[12] Third Generation Partnership Project : proyecto conjunto de tercera generación de comunicaciones móviles, con grandes facilidades para la conexión a los servicios de Internet

[13] Quality of Service: calidad del servicio

[14] Movimiento irregular ligero, variación o inestabilidad especialmente en señales eléctricas o en dispositivos electrónicos.

- Eco: ocurre cuando la voz propia o de otra persona cercana se “suma” a la voz escuchada en una comunicación de voz.

A continuación, se explicarán más detalladamente dichos factores y se analizará posteriormente como influyen cada uno de ellos en el procesamiento automático de la voz, específicamente en el reconocimiento de locutores en VoIP.

3.1 Demora punto-punto

Esta demora es la suma de todas las demoras en un sentido que ocurren en una llamada. Según la recomendación G 114 (ITU-T, 2003) una demora entre 150 y 400 ms. se acepta sin distorsión apreciable. Dicha demora está determinada por cuatro causas:

- Propagación: es el tiempo tomado por la voz para propagarse de un punto a otro en la red, induciendo una demora en un sentido entre 70 y 100 ms., lo cual es imperceptible para el oído humano. Dicha demora se ve afectada por dos causas adicionales, la demora de empaquetamiento y la demora propia de la propagación de los paquetes a través de la red, lo que la hace variable durante la trasmisión.
- Compresión: para aprovechar mejor el ancho de banda del canal de comunicación que soporta la red, se aplican técnicas de compresión por codificación de la voz en los paquetes. Dicha compresión introduce demoras que dependen del tipo de códec utilizado.
- Empaquetamiento: es el tiempo requerido para llenar cada paquete con las tramas de rasgos de voz. En la mayoría de los casos, mientras mayor sea el paquete es mayor el tiempo requerido de empaquetamiento. Los códec controlan esta causa de demora.
- Conmutación de paquetes: es el tiempo requerido por los interfaces de conmutación y enrutamiento para colocar un paquete en los buffer de almacenamiento y tomar la decisión hacia cual nuevo interfaz debe dirigirlo. Esta es la razón por la cual la arquitectura de dichos interfaces, así como las de sus buffer de almacenamiento, son factores críticos de diseño de la red para reducir dicha demora al mínimo.

3.2 Demora por “jitter”

En VoIP algunas demoras son relativamente constantes como las explicadas en el epígrafe anterior pero otras dependen de las condiciones de la red. Esta demora es la variación inestable en el tiempo de arribo de los paquetes en el receptor lo cual afecta, mucho más que la latencia, la calidad percibida de una conversación.

El extremo transmisor envía los paquetes a intervalos regulares, dicha velocidad de trasmisión se afecta por las congestiones y por una inapropiada cola en los buffer de almacenamiento ya que los paquetes correspondientes a una misma conversación recorren diferentes caminos y al llegar al extremo receptor traen diferentes demoras no controladas. El “jitter” afecta la voz en forma tal que si los paquetes consecutivos no arriban en el momento adecuado al extremo receptor, puede provocar que la voz sintetizada no sea continua, siendo perceptible para el oído humano. El “jitter” debe mantenerse entre 30 y 75 ms. y depende del tamaño del paquete y de los buffer de almacenamiento así como del códec utilizado [3]

3.3 Razón de entrega de paquetes

El protocolo TCP-IP no puede asegurar que lleguen todos los paquetes a su destino y mucho menos en el orden en que fueron trasmitidos, pudiendo incluso perderse paquetes por fallos de los dispositivos de

enrutamiento o congestiones en la red. Para logar una calidad adecuada en la voz, la razón de entrega de paquetes en el extremo receptor debe ser superior a un 99% [3]

En ocasiones, debido a la latencia o el jitter, los paquetes llegan tarde y se congestionan los buffers de almacenamiento, resultando en la pérdida de paquetes, provocando discontinuidades molestas en la voz sintetizada en el extremo receptor. Los métodos de retrasmisión de paquetes no siempre son efectivos ya que la demora en la llegada del paquete retrasmitido puede no hacerlo ya útil para conservar la continuidad de la conversación. La pérdida de paquetes en burst, que puede ocurrir frecuentemente en redes inalámbricas, afecta seriamente la comunicación. En dependencia del códec utilizado, múltiples tramas de voz pueden venir en un paquete, y su pérdida puede afectar seriamente la voz.

La pérdida de paquetes es inevitable y se utilizan técnicas para ocultar las mismas conocidas como PLC[15]. Cuando ocurre una pérdida de paquetes los sistemas tratan de ocultarlo de variadas formas, la más común es la de remplazar los paquetes perdidos, repitiendo el paquete perdido o repitiendo el ultimo recibido. Puede realizarse también una interpolación entre el paquete previo y el que sigue, que consiste en llenar la primera mitad del paquete perdido con el paquete anterior y la segunda mitad con el paquete posterior, este método es común aplicarlo a las ráfagas de paquetes perdidos, interpolando con el anterior y el posterior a los paquetes perdidos[16].

3.4 Compresión de la voz

Desde la aparición de las comunicaciones digitales, el aprovechamiento eficiente del ancho de banda del soporte de comunicación ha sido siempre un objetivo a alcanzar, mientras menos ancho de banda requiera un servicio, la red puede atender más usuarios y brindar más servicios [3]. En este contexto, utilizar el menor ancho de banda posible manteniendo una calidad aceptable para los usuarios, representa un reto.

Los métodos de compresión de la voz o códec de voz permiten que los sistemas trasmitan la voz analógica sobre las redes digitales. En VoIP el trasmisor aplica una codificación a la voz para comprimir la cantidad de información que se envía a través de la red, la voz viaja por la red en forma codificada y debe ser decodificada y re-sintetizada en el extremo receptor para obtener una voz "similar" a la trasmitida.

Los códec se clasifican según la frecuencia a que se digitaliza la voz ("fm": frecuencia de muestreo) en banda estrecha ("NB: narrowband", fm=8 kHz) o banda ancha ("WB: wideband", fm=16 kHz) y según la razón de bits de la codificación, como de bajo o alto "bit rate"[17].

En la figura 2 se muestran los anchos de banda ocupados por ambos códec. Mientras más calidad se requiera de la voz, mayor ancho de banda requerirá el códec, convirtiéndose en un problema lograr una adecuada eficiencia en el uso de la red.

---

[15] Packet Loss Concealment: Ocultamiento de la perdida de paquetes.

[16] Un estudio sobre los diversos métodos PLC puede encontrarse en: "Comparative Study of Techniques to minimize packet loss in VoIP", de Shveni P. Mehta, 21st Computer Science Seminar, 2005.

[17] Razón o velocidad de bits, se mide en kilobits por segundo: kb/s

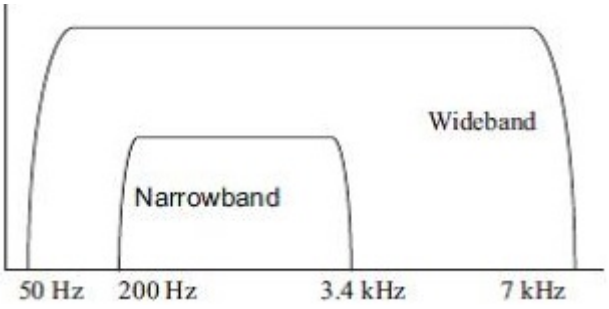


Fig. 2. Anchos de banda de la voz en códec de banda estrecha y de banda ancha [3].

Se establece una llamada entre dos puntos utilizando un protocolo de señalización para que ambos puntos se "pongan de acuerdo" en cual tipo de códec utilizaran, en VoIP los usuarios de ambos extremos tienen una lista de códec ya que las redes pueden manejar diferentes niveles de compresión de la voz en dependencia del servicio y del QoS requerido [1]. Esta facilidad trae como consecuencia el problema de la trans-codificación: una comunicación VoIP en su recorrido por la red, puede codificarse y decodificarse por diferentes códec, en dependencia del ancho de banda y la QoS requerida por el soporte de comunicación por donde se esté trasmitiendo. La trans- codificación implica afectaciones en la calidad de la voz y en la demora punto-punto.

La tabla del anexo 1 refleja las principales características técnicas de los códec estándar utilizados. Los códec de banda estrecha "NB" G.711 (64 kb/s), G.726 (16, 24, 32 kb/s), G.728 (16 kb/s), G.729 (6.4, 8, 11.8 kb/s) y G.723.1 (5.3, 6.3 kb/s) son ampliamente utilizados en comunicaciones de VoIP. Observe que los códec G.711 y G.726, aunque son de alto bit-rate son considerados de banda estrecha, ya que el método de codificación utilizado (PCM) ocupa solo dicha banda.

Por otra parte los códec de banda ancha "WB" como G.722 (48, 56, 64 kb/s) y G.722.1 (24, 32 kb/s), mejoran la inteligibilidad y naturalidad del habla, pero requieren un mayor "bit rate" [1].

La solución para aprovechar el máximo ancho de banda de la red, está en los códec adaptativos, que ajustan su bit rate en función de las condiciones de la red en cada momento [4]. El códec AMR-WB[18] ha sido estandarizado por 3GPP y por ITU-T para aplicaciones de habla conversacional, lo cual es muy significativo porque, por primera vez un códec ha sido adoptado por VoIp en comunicaciones alambradas o inalámbricas, indistintamente. Esto elimina la necesidad de la trans-codificación y facilita la implementación de aplicaciones y servicios de voz de banda ancha a través de un amplio rango de plataformas y sistemas de comunicación. Por otra parte, otro códec adaptivo, el AMR-NB[19] ha sido estandarizado por 3GPP como códec para los sistemas móviles de tercera generación.

Otros estándares para aplicaciones específicas de VoIP son el GSM en sus tres variantes de bit-rate: HR (5.6 kb/s), FR (13 kb/s) y EFR (12.2 kb/s), utilizado en la telefonía móvil, el iLBC[20] (13.33 y 15.2 kb/s) y el iSAC[21] (10 a 32 kb/s) para aplicaciones específicas de VoIP entre computadoras como Skype y GoogleTalk.

Por último se reportan códec de tipo multimodal con versiones de banda estrecha (Speex-NB: 2.15 a 24.6 kb/s) y BroadVoice: 16 kb/s) y de banda ancha (Speex-WB: 4 a 44.2 kb/s y BroadVoice: 32 kb/s).

[18] Adaptive Multi-Rate Wideband: códec adaptivo multi-velocidad de banda ancha.

[19] Adaptive Multi-Rate Narrowband: códec adaptivo multi-velocidad de banda estrecha.

[20] Internet Low Bitrate Codec: códec de baja velocidad para Internet.

[21] Internet Speech Audio Codec: códec de audioy habla para Internet.

3.5 El eco en la comunicación de voz

Existen tres tipos de ecos que impactan en las comunicaciones de VoIP: el FEXT[22], el NEXT[23] y el efecto combinado de ambos [2].

El FEXT es causado por la conversión hibrida[24] ocurrida en la planta telefónica, el usuario oye su propia voz que "rebota" en la hibrida con una cierta demora.

El NEXT ocurre cuando se utilizan micrófonos y altoparlantes para establecer la llamada o hay dos personas hablando en un mismo local, este es un caso muy común cuando se realizan llamadas entre computadoras usando VoIp (chat de voz, Skype, GoogleTalk, etc.). La voz emitida por los altoparlantes es captada por los micrófonos y devuelta hacia el usuario remoto. Este efecto de eco puede considerarse un ruido adicionado a la voz del locutor que se va a reconocer y es conocido como "babble" o ruido de murmullo, el cual es muy difícil de reducir o compensar debido a que posee las mismas características de la voz que debe ser reconocida. Las técnicas de "Blind Source Separation"[25] pueden ser utilizadas para separar el eco de la señal de voz de interés [2].

3.6 La QoS de la VoIP sobre redes inalámbricas

Las redes inalámbricas de área local WLAN son una de las tecnologías de soportes de comunicación más desarrolladas en la actualidad, ofreciendo movilidad para el acceso a las computadoras y mayor flexibilidad a los "IP-phone", jugando una importante función en las redes de voz de nueva generación. Tienen su acceso no alambrado a la red LAN por radiofrecuencia o infrarrojo y utilizan un protocolo diferente para su acceso, caracterizándose por su movilidad, simplicidad, escalabilidad y bajo costo [4].

La comunicación entre dos usuarios de una WLAN se realiza comúnmente a través de un AP[26] utilizando el protocolo estándar de la IEEE 802.11, en diferentes bandas de frecuencia y "bit rate". Las WLAN proveen conexión a las redes TCP/IP, convergiendo ambas tecnologías, lo que permite la incorporación de la VoIP sobre las WLAN, conociéndose como VoWLAN[27].

La QoS de la VoWLAN presenta aún mayores restricciones que las de VoIP, porque la demora por jitter y la pérdida de paquetes son significativamente mayores en una WLAN. Son restricciones específicas de la QoS sobre VoWLAN las siguientes [5]:

- Aparecen otros factores propios de la comunicación por radiofrecuencia como la pérdida de cobertura y la interferencia con otras fuentes de radio como teléfonos inalámbricos, dispositivos "bluetooth" u hornos de microondas, debido a que se utiliza una banda de frecuencias de radio compartida con dichas tecnologías. Dicha degradación de la calidad de la comunicación y su consecuente reducción en el ancho de banda máximo a utilizar, provoca una nueva afectación en la QoS, propia de la VoWLAN.

---

[22] Far End Crosstalk: diafonía o eco lejano. La diafonía es la transferencia de señales no deseadas entre canales de comunicación.

[23] Near End Crosstalk: diafonía o eco cercano.

[24] Dispositivo electrónico en la planta telefónica que convierte el par telefónico del usuario en dos pares telefónicos, para separar las voces de la llamada. Este fenómeno es común también en las comunicaciones satelitales de larga distancia pero por otras razones.

[25] Separación ciega por la fuente: técnica de análisis de componentes de una mezcla de fuentes, que trata de separar las fuentes (o señales) mezcladas sin conocer las fuentes ni la forma en que se mezclaron.

[26] Access Point: punto de acceso de una red inalámbrica

[27] Voice over wirelees LAN: voz sobre redes locales inalámbricas

- Cuando hay varios usuarios conectados al mismo AP, se provocan congestiones implicando mayores demoras punto a punto y por jitter, al requerirse retrasmisiones que además elevan la razón de pérdidas de paquetes.
- Como el usuario puede moverse entre los AP, se pueden introducir demoras de conmutación entre AP de hasta 500 mseg., que son claramente audibles, sobre todo cuando el usuario se mueve entre dos AP de diferentes sub-redes, este fenómeno trata de evitarse diseñando la red con la menor probabilidad de conmutación entre sub-redes.

4 Análisis de la influencia de la QoS de la VoIP sobre los métodos de reconocimiento del locutor

Los actuales métodos de procesamiento automático del habla, específicamente el reconocimiento del locutor, utilizan métodos de extracción de parámetros y de clasificación cuyo principal objetivo es el de obtener de una expresión de habla, la mayor cantidad de información que identifique al locutor que la expresó, libre de otras informaciones propias del contenido lingüístico, del comportamiento del canal así como de otros fenómenos. La variabilidad en el habla, tanto intrínseca (propia de la persona, debido al entorno comunicativo, al estado emocional y de salud, etc.) como la extrínseca (propia del canal, el ruido, etc.) constituye el principal reto a enfrentar para lograr una mejor eficacia en el reconocimiento del locutor.

Las restricciones de QoS propias del protocolo TCP/IP y que por tanto están presentes en la VoIP, provocan fenómenos indeseables en el habla procesada y trasmitida por Internet, que le introducen variabilidad extrínseca, propia de la VoIP. Dichos fenómenos se observan en la perdida de información debido a la compresión espectral y las demoras, discontinuidades espectro-temporales con la consiguiente distorsión espectral, etc., que modifican la información que identifica al locutor, afectando la eficacia del método de reconocimiento.

A continuación se expondrá como las restricciones de la QoS de la VoIP influyen en la eficacia del reconocimiento del locutor, a través del análisis de resultados reportados en la última década.

4.1 Influencia de la pérdida de paquetes en el reconocimiento del locutor

Para llevar a cabo un experimento que permita evaluar la influencia de la perdida de paquetes en el reconocimiento del locutor, se requiere simular la red con un modelo sencillo. Dicha simulación, aunque garantiza un control total de las condiciones experimentales, no garantiza una confiabilidad máxima en los resultados; de lo contrario durante la realización de experimentos en condiciones reales se dificulta el control de los parámetros de trasmisión y la repetición de ciertos experimentos en las mismas condiciones [6].

4.1.1 Modelo de Gilbert para simular la pérdida de paquetes en VoIP

Los estudios sobre las distribuciones de las pérdidas de paquetes en Internet [6] han demostrado que este proceso puede caracterizarse usando modelos de Markov.

Un modelo de Markov de dos estados conocido como modelo de Gilbert [9] puede capturar la dependencia entre esas pérdidas temporales y se muestra en la figura 3.

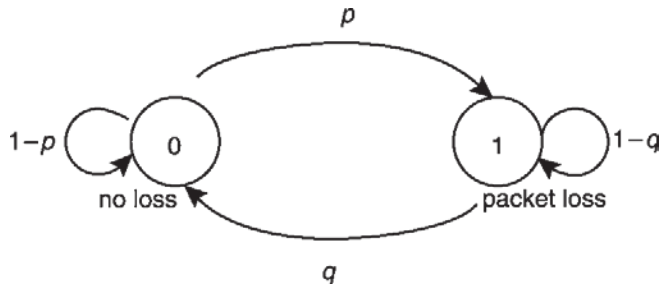


Fig. 3. Modelo de Gilbert para simular la pérdida de paquetes en las redes TCP/IP [9].

El estado 1 representa una pérdida de paquete y el estado 0 representa que no hay pérdida, las probabilidades de transición son *p* y *q*, *p* es la probabilidad del estado 0 al 1 (perder un paquete si el anterior no se perdió) y *q* es la probabilidad del estado 1 al 0 (recibir un paquete después de uno perdido), (1-*q*) es la probabilidad condicional de pérdida. Diferentes valores de *p* y *q* definen diferentes condiciones de pérdida de paquetes.

La probabilidad que *n* paquetes consecutivos se pierdan está dada por $p(1-q)^{n-1}$. Si $(1-q) > p$, entonces la probabilidad de perder un paquete después de haber perdido el anterior es mayor que la probabilidad de haber perdido un paquete después de haber recibido el anterior, lo cual es muy común en TCP/IP donde ocurren pérdidas de paquetes por ráfagas. Observe que $p+q$ no es necesariamente 1.

En la tabla 1 se recogen probabilidades asignadas a *p y q* en el modelo de Gilbert, que han sido reportadas para diferentes condiciones de la red:

Tabla 1. Probabilidades asignadas a *p* y *q* en diferentes reportes.

Condición de la red	***p***	***q***	**Referencia**
promedio	0.1	0.7	[6],[7],[8]
	0.057	0.94	[6]
mala	0.25	0.4	[7],[8]

4.1.2 Influencia de los métodos para ocultar la pérdida de paquetes en el reconocimiento de locutores

Los métodos de PLC están muy asociados al tipo de códec, por ejemplo entre los códec de banda estrecha, G.721, G.723, G.728, G.729 y AMR no utilizan PLC, sin embargo, G.711, AMR-NB, ILBC-N y Speex-NB si la utilizan [10].

Un análisis del efecto del PLC sobre la eficacia de la verificación del locutor independiente del texto, medida en perdida relativa de EER[28], en los códec de banda estrecha G.711(64 kb/s), ILBC-N (15.2 kb/s), Speex-NB (15.0 kb/s) y SILK (15 kb/s, códec utilizado en Skype) fue llevado a cabo en el 2011 [10], para pérdidas de paquetes de 5, 10, 15 y hasta 20 % , seleccionados de forma aleatoria. Se tomó como referencia el EER sin pérdida de paquetes, los resultados de pérdida relativa del EER se muestran en la figura 4.

[28] Equal Error Rate: razón de igual error, medida de eficacia para verificación de locutor obtenida de la curva de comportamiento del error de detección.

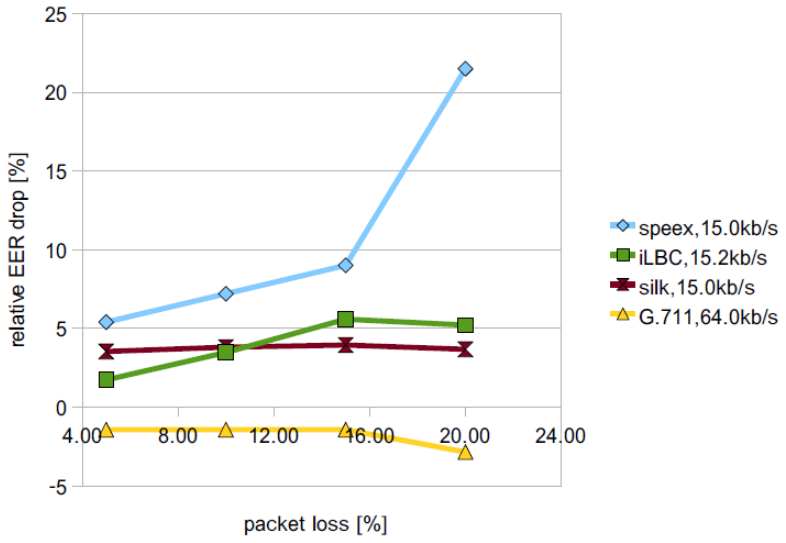


Fig. 4. Efecto del PLC para diferentes códec [10].

Los resultados muestran que el PLC en el códec G.711 mantiene la eficacia incluso hasta un 20% de pérdida de paquetes, los codec iLBC y Speex sufren una ligera degradación, no sucediendo así con el SILK, donde el PLC no logra mantener la eficacia. Se comprueba que el bajo nivel de compresión del códec G.711 (64 kb/s) favorece el comportamiento del PLC en el mantenimiento del EER incluso para altos % de pérdida de paquetes, lo que no ocurre con los otros códec evaluados que poseen un mayor nivel de compresión con un correspondiente bajo "bit rate" (15 kb/s) donde el PLC no puede evitar que ocurra cierta degradación en el EER, que llega a ser critica para el códec Speex en el 20 % de pérdida de paquetes.

4.1.3 Influencia de la pérdida de paquetes en el reconocimiento de locutores

Aparecen reportados varios experimentos de reconocimiento de locutores, utilizando el modelo de Gilbert [9] para simular la pérdida de paquetes:

- En el 2005 [6] se evalúan tres condiciones de la red, dos simulando una red promedio (ver tabla 1) y una tercera condición real (pérdida promedio de paquetes: 5.26%), utilizando un códec G.711, con una duración de paquetes de 20 ms. Se comparan los resultados de identificación de 10 locutores con los de una red sin códec ni pérdida de paquetes. Se extraen rasgos LPC[29] de la voz.

  Se observa la degradación en la eficacia, para las tres condiciones de la red, aunque los resultados con la condición real son los más cercanos a los obtenidos sin códec ni pérdida de paquetes, lo que indica la no confiabilidad de los experimentos simulados.
- En el 2004 [7] se evalúan dos condiciones de la red, simulando una red promedio y una red mala (ver tabla 1) así como una red ideal, sin pérdida de paquetes, utilizando dos códec, uno de alto bit-rate G.711 y uno de bajo bit-rate G.723.1, comparándolos con una red sin códec. Se realiza un experimento de reconocimiento de locutores independiente del texto, se extraen rasgos LFCC[30] más

[29] Linear Prediction Coefficient: coeficientes de predicción lineal del filtro que modela el tracto vocal
[30] Linear Frequency Cepstral Coefficient : coeficientes cepstrales en escala lineal de frecuencias

sus derivadas modelándose cada muestra de voz con la adaptación GMM[31]-UBM[32]-MAP[33], la decisión se obtiene usando el criterio LLR[34]. Los resultados de EER del reconocimiento del locutor se muestran en la tabla 2:

Tabla 2. Resultados de reconocimiento de locutores en EER, con/sin pérdida de paquetes.

Condición de la red	**Tipo de códec**		
	Sin códec (128 kb/s)	**G.711 (64 kb/s)**	**G.723.1 (5.3 kb/s)**
sin perdida	0.25%	0.25%	2.68%
promedio	0.25%	0.25%	6.28%
mala	0.5%	0.75%	9%

Se observa que, independiente de la condición de la red, el EER se mantiene bajo cuando no se utiliza códec, o cuando la compresión es muy baja (G.711) no sucediendo así cuando se utiliza un alto nivel de compresión (G.723.1). Este resultado demuestra que en sentido general, la pérdida de paquetes no provoca una afectación apreciable en el EER del reconocimiento de locutores con independencia del texto, a menos que la VoIP se codifique con alta compresión (bajo "bit rate").

- Un experimento adicional reportado en [7], fue realizado para confirmar qué cantidad de información puede perderse en los paquetes, sin degradar apreciablemente el comportamiento del reconocimiento de locutores con independencia del texto. Consistió en dos pruebas con secuencias de 4 dígitos y de un digito, entrenando cada locutor con 30 segundos de su voz. Se extraen rasgos MFCC[35] más sus derivadas, modelándose con GMM -UBM -MAP y usando criterio de decisión LLR. Para simular la perdida de paquetes, se procedió a eliminar de las secuencias de pruebas diferentes % de tramas de vectores de rasgos y evaluar el correspondiente EER. Los resultados de EER con relación al % de vectores de rasgos eliminados (perdidos) se muestran en la figura 5.

  Se comprobó que con la pérdida de hasta un 75 % de los vectores de rasgos, el error prácticamente no se incrementa. Confirmándose que en los vectores de rasgos existe suficiente redundancia que identifica a la persona, como para reducirlos a un cuarto del total sin prácticamente afectar la eficacia.

[31] Gaussian mixture models: modelo de mezclas gaussianas que representa una muestra de voz
[32] Universal background model: modelo GMM de muestras de voces de una población
[33] Maximum a posteriori : método de adaptación del GMM al UBM
[34] Log-likelihood ratio: razón de similaridad logarítmica que es el resultado de comparar dos muestras de voz
[35] Mel Frequency Cepstral coefficient : coeficientes cepstrales en escala de frecuencias mel (escala percepción auditiva)

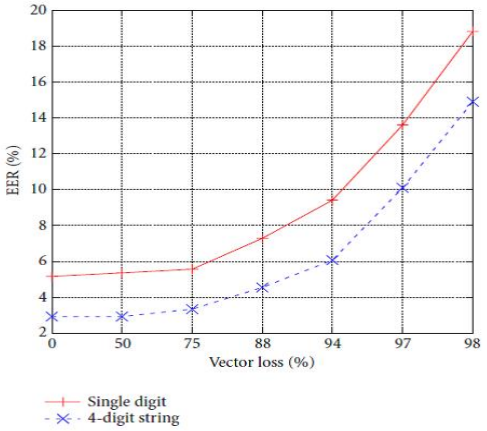


Fig. 5. Eficacia del reconocimiento de locutor en EER con relación al % de vectores de rasgos perdidos [7].

- También en el 2004 [11], se llevó a cabo un experimento que relaciona el tamaño del paquete (en cantidad de tramas de rasgos) con la razón de pérdida de paquetes, en identificación de locutores, demostrándose que a medida que se reduzca el tamaño del paquete se empeora la eficacia del reconocimiento. Se utilizaron modelos GMM con 10 componentes para modelar la voz. Se utiliza como medida de comparación el % de identificaciones acertadas. En la figura 6 se observan los resultados.

  Se observa una eficacia de identificación superior al 90% cuando el paquete tiene 32 o más tramas de rasgos, empeorando apreciablemente la eficacia a medida que se reduce la cantidad de tramas por paquete a 16 o menos siendo peor el comportamiento por supuesto, con un mayor % de pérdidas de paquetes.

  Se propone un método de entrenamiento de los modelos de los locutores que tiene en cuenta el % de pérdida de paquetes que sufre la voz del locutor desconocido, que eleva la eficacia de la identificación a más del 90%.

  Se comprobó además que un incremento de la duración de la expresión de prueba favorece un incremento de la eficacia de identificación a niveles superiores al 90%, al contar con más muestras para comparar, independiente del tamaño del paquete y del % de pérdida de paquetes.

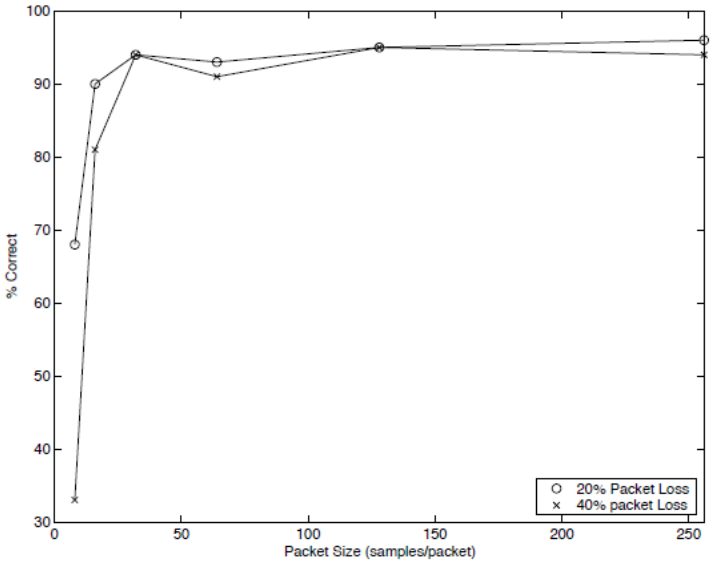


Fig. 6. Rendimiento de identificación del locutor en % vs. tamaño del paquete [11].

4.2 Influencia del ancho de banda y el "bit rate" del códec en el reconocimiento de locutores

El reconocimiento de locutores sobre VoIP debe enfrentar además el reto de la compresión por medio de dispositivos códec para codificación y decodificación de la voz, durante su recorrido entre el extremo trasmisor y el receptor. En el peor escenario y en dependencia de la configuración de la red, su ancho de banda y la QoS requerida, la voz puede trans-codificarse, o sea, pasar por diferentes códec durante su recorrido.

La figura 7 esquematiza dicho proceso. Tanto en el proceso de codificación, que lleva a cabo una compactación de la voz original (pto. 1) para reducir el bit-rate en que se trasmite (pto.4), como en su posterior decodificación (pto. 3) para re-sintetizarla de nuevo (pto.2), se provocan pérdidas de la información contenida en la voz, incluida información de identidad de la persona, afectando la eficacia del reconocimiento de locutores, por lo que se requiere obtener métodos de reconocimiento que sean robustos al paso de la voz por los códec.

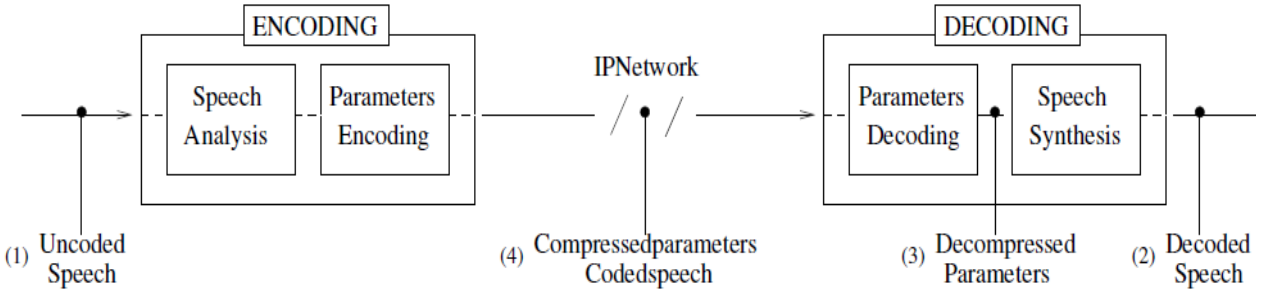


Fig. 7. Esquema del proceso de compresión de la red sobre la voz [12].

4.2.1 Influencia de la compresión del códec en el reconocimiento del locutor

- En un trabajo publicado en el 2011 [13], se utiliza un clasificador GMM-SVM[36] de 256 mezclas y utilizando rasgos MFCC y, para evaluar el comportamiento de la identificación de locutores ante la codificación de la voz por diferentes códec. En dicho estudio se evalúan dos condiciones de desigualdad (del inglés: "missmatch") entre entrenamiento y prueba:
  - Condición "pareja": donde la voz para entrenamiento y prueba se codifica con el mismo códec, o no se codifica en lo absoluto.
  - Condición "no pareja": donde la voz para entrenamiento y prueba se codifica con diferentes códec, o una de las dos no se codifica.

  Ambas condiciones de desigualdad son bastante comunes en la VoIP, ya que las configuraciones de las redes y los interfaces con las mismas, pueden cambiar en el tiempo para un mismo usuario y no son conocidas por este. Un usuario puede acceder a la red en un momento a través de un "ip-phone" y en otro momento usar un dispositivo móvil o utilizar un servicio de voz, cada interface utiliza diferentes códec.

  Se evaluaron seis códec de banda estrecha muy utilizados en telefonía fija, móvil y VoIP: G.711 (64 kb/s), G.723.1 (6.4 kb/s), G.729 (8 kb/s), GSM FR 06.10 (13 kb/s), GSM EFR 06.60 (12.2 kb/s), y Speex-NB (8 kb/s). La tabla 3 muestra los resultados de eficacia medida en % de precisión para las diferentes combinaciones de códec, la diagonal de la tabla es la condición pareja:

Tabla 3. Precisión en la identificación de locutores para diferentes códec [13].

training/testing	un-coded	G.711	G.723	G.729	GSM 06.10	GSM 06.60	Speex 8	average	stddev
uncoded	**89.67**	87.40	49.49	51.49	51.49	51.44	83.63	66.37	17.59
G.711	86.42	**88.23**	46.98	47.02	50.47	64.65	80.60	66.34	16.07
G.723.1	71.63	68.88	**73.81**	61.63	60.33	71.30	75.58	69.02	4.64
G.729	65.16	62.37	57.95	**77.12**	37.72	77.91	62.74	63.00	8.91
GSM 06.10	69.63	70.98	55.26	44.98	**83.12**	57.72	72.23	64.84	10.45
GSM 06.60	72.00	65.54	63.07	63.21	41.86	**84.28**	63.07	64.72	7.90
Speex 8	86.60	84.23	62.88	53.07	62.79	67.67	**86.65**	71.99	11.87

El comportamiento de la identificación de locutores en condición pareja, sigue el comportamiento de la calidad del habla medida para dicho códec que en orden decreciente es: G.711, Speex, GSM EFR, GSM FR, G.729 y G.723.1 Los mejores resultados en condición pareja son para la voz no codificada ("uncodec") y le siguen los códec G.711, Speex y los GSM.

En condición no pareja se observa una peor precisión en todos los casos, pero el decrecimiento es peor mientras más no parejos estén los códec del entrenamiento y la prueba. Cuando el entrenamiento se hace sin códec o con un códec de calidad (G.711 o Speex) y la prueba se hace sin códec o con un códec de calidad, la precisión no disminuye tanto, respecto a la condición pareja. Sin embargo si se prueba con los códec de peor calidad la precisión llega a disminuir hasta casi la mitad de la condición pareja.

- En un trabajo del 2012 de los mismos autores [14] pero con experimentos de verificación de locutores con modelos GMM-UBM-MAP y utilizando los mismos códec, llegan a conclusiones similares respecto al comportamiento de la eficacia medida en EER.

[36] Gaussian Mixture Models- Support Vector Machine: método de clasificación de reconocimiento de locutores que concatena en un vector las medias de las GMM y las clasifica con máquinas de soporte vectorial.

Estos resultados confirman que en el reconocimiento de locutores sobre VoIP:

- La calidad del códec, medida no solo por el método de codificación, sino por el grado de compresión aplicada a la voz, dada por el "bit-rate", influye en la eficacia.
- La condición pareja de los códec para la voz del entrenamiento y de la prueba es la condición ideal, cualquier otra condición no pareja (lo cual es muy común y no fácil de determinar) afecta la eficacia.
- Llevar acabo el entrenamiento con una voz codificada con un códec de mejor calidad (G.711 o Speex) asegura los mejores resultados de eficacia para cualquier otra condición no pareja.

- En otro trabajo muy completo del 2011 [10], se utiliza un clasificador GMM-UBM-MAP con 512 mezclas y rasgos MFCC con sus derivadas, para evaluar el comportamiento de la eficacia en verificación de locutores ante la codificación de la voz en condiciones pareja y no pareja de desigualdad y se compara con la voz codificada con G.711 (64 kb/s). La condición no pareja se entrena con la voz sin aplicarle códec.

  Se evaluaron diez códec de banda estrecha muy utilizados en telefonía fija, móvil y VoIP: G.723.1 (5.3 y 6.3 kb/s), G.726 (16, 24 y 32 kb/s), G.728 (16 kb/s), G.729 (6.4, 8 y 11.8 kb/s), GSM FR (13 kb/s), GSM HR (5.6 kb/s), AMR-NB (4.8 a 11.2 kb/s), iLBC (13.3 y 15.2 kb/s), Speex-NB (4, 8 y 15 kb/s) y SILK (5, 8 y 15 kb/s).

  En la figura 8 se observa una comparación para la condición no pareja, de la eficacia de la verificación de locutores medida en EER, respecto al "bit rate" de cada códec, para las 28 combinaciones de códec evaluadas.

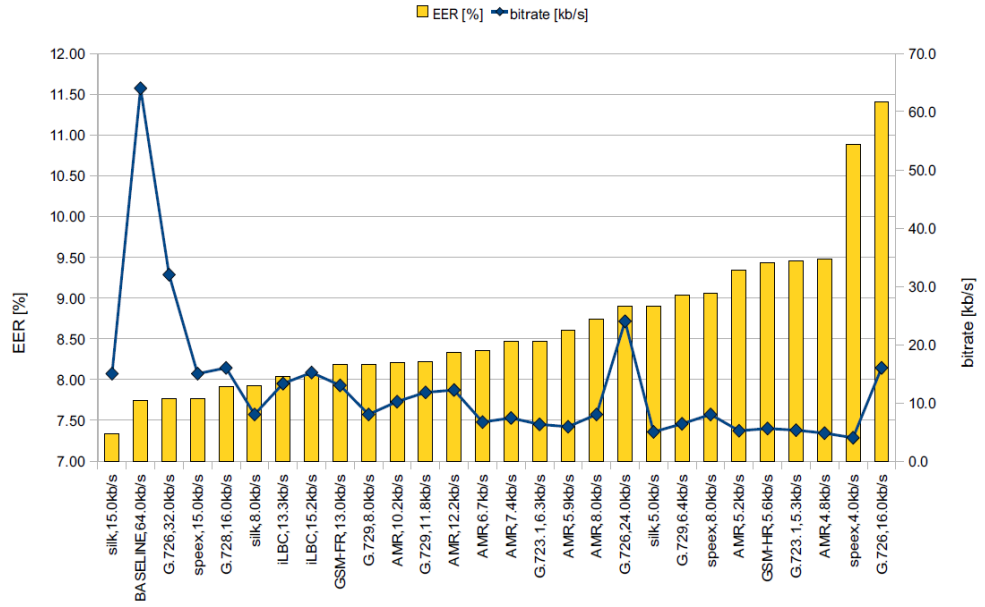


Fig. 8. Eficacia de verificación (EER) y bit rate (kb/s) de las 28 combinaciones de códec [10].

De dicha figura se confirma que:

- El incremento de la compresión de la voz en búsqueda de un mejor aprovechamiento del ancho de banda del canal de comunicación, con la reducción del "bit rate" en los códec, provoca una afectación en la eficacia de la verificación del locutor que se incrementa a medida que el "bit rate" disminuye.

- Para la condición no pareja, el códec SILK (8 y 15 kb/s) presenta el mejor comportamiento de eficacia conservando la información identificativa del locutor de forma muy similar al G.711 (64 kb/s) y el G.726 (32 kb/s) que poseen un "bit rate" superior. Menos de un 5% de pérdida de eficacia se observa además en el G.728 (16 kb/s), iLBC (13.3 y 15.2 kb/s), Speex (15 kb/s) y el mismo G.726 (32 kb/s). De forma opuesta el comportamiento del G.726 (16 y 24 kb/s) es muy malo al compararse con otros códec con semejantes "bit rate".
- Para la condición pareja, se observa un incremento de la eficacia en las versiones de alto "bit rate" de los códec G.726 (32 kb/s), G.729 (11.8 kb/s), AMR-NB (8, 10.2 y 12.2 kb/s), iLBC (15.2) y SILK (8 y 15 kb/s). Sin embargo esos mismos códec para bajo "bit rate" presentan mayor degradación que para la condición no pareja, lo que puede deberse al hecho del bajo "bit rate" que provoca una reducción de la información identificativa en la voz del locutor, no obstante tratarse de una condición pareja entre entrenamiento y prueba para el reconocimiento de locutores.

- Otros autores en trabajos publicados en 2000 [15] y el 2008[8] evalúan la influencia del códec GSM, utilizado en telefonía celular, en sus tres variantes de calidad GSM FR 06.10 (13 kbit/s), GSM HR 06.20 (5.6 kbit/s) y GSM EFR 06.60 (12.2 kbit/s), en la identificación [15] y verificación [8] de locutores. Se utiliza la base de datos TIMIT con frecuencias de muestreo a 16 kHz y a 8 kHz y se extraen coeficientes cepstrales, clasificando con GMM de 16 mezclas.

  En la tabla 4 se muestran los resultados de eficacia de identificación medida en % de error y la eficacia de verificación medida en % de EER, observándose de nuevo una degradación de la eficacia correspondiente con la calidad del códec utilizado, el códec que más comprime la voz (GSM HR 06.20) es el que peor eficacia brinda. Puede observarse además, que la reducción en la frecuencia de muestreo de la base, ya implica una perdida apreciable de eficacia.

Tabla 4. Error de identificación y de verificación sin códec y para diferentes códec GSM [15] y [8].

	Sin códec		**GSM**		
	TIMIT 16 kHz	**TIMIT 8 kHz**	**FR**	**HR**	**EFR**
Identificación	2.2	13.1	31.5	38.5	28.2
Verificación	1.1	5.1	7.3	7.8	6.6

Estos resultados confirman que en el reconocimiento de locutores sobre VoIP:

- La eficacia de la verificación del locutor se afecta más a medida que el códec comprime más la voz y se reduce el "bit-rate".
- La condición pareja entrenamiento-prueba eleva la eficacia del reconocedor con respecto a la condición no pareja, pero solo para altos "bit-rate" del códec.

4.2.2 Influencia del ancho de banda de los códec en el reconocimiento del locutor

La VoIP requiere la utilización de códec que, según la frecuencia a que se muestree la voz para ser codificada, son de banda estrecha (NB) o de banda ancha (WB)[37] (ver epígrafe 3.4 y figura 2).

Se ha demostrado que la extensión de banda de NB a WB contribuye en cierta medida a una mejor inteligibilidad y calidad percibida y una mayor eficacia en el reconocimiento del locutor, tanto por vía forense, como automático. Las señales WB contienen frecuencias adicionales que potencialmente aumentan la información identificativa del locutor, las bajas frecuencias (entre 50 y 200 Hz) incluyen la frecuencia fundamental[38] y en ocasiones el primer formante[39], mientras que las altas frecuencias (entre

[37] Recientemente se ha estandarizado una nueva banda "SWB: super wideband", desde 50 hasta 14000 Hz, con frecuencia de muestreo fm=32 kHz, utilizada para videoconferencias de alta calidad [16].

[38] Frecuencia de resonancia de las cuerdas vocales, conocida también por "pitch"

3400 y 7000 Hz) incluyen los formantes más altos (cuarto y quinto) y otras características de sonidos nasales y fricativos que son también identificativas [17].

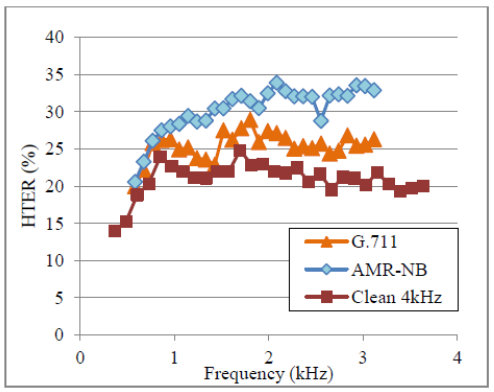
Se conoce además que la información identificativa de un locutor contenida en la voz, no está distribuida uniformemente en todas las bandas de frecuencias, siendo unas más discriminativas que otras[40]. Por ejemplo la frecuencia fundamental y los formantes de las vocales se manifiestan en las mujeres a más altas frecuencias que en los hombres al tener un tracto vocal más pequeño y las consonantes nasales son identificativas en bajas y altas frecuencias [18].

- Un trabajo del año 2014, [16] lleva a cabo un análisis de la perdida de la información identificativa del locutor en diferentes bandas espectrales, para diferentes códec. El experimento se lleva a cabo comprimiendo la voz con dos códec NB, el G.711 (64 kb/s) y el AMR-NB (12.2 kb/s) y con dos códec WB, el G.722 (64 kb/s) y el AMR-WB (12.65 kb/s). Adicionalmente se utiliza la voz sin comprimir con frecuencias de muestro NB (fm = 8 kHz) y WB (fm = 16 kHz).

  Para las seis condiciones de compresión se divide el espectro de la voz con 32 filtros que se agrupan en 28 sub-bandas, extrayendo 4 coeficientes LFCC a cada una de las salidas de cada sub-banda. Con esos vectores de rasgos por sub-banda y por condición se realizan 28 x 6= 168 experimentos de verificación de locutores utilizando el método de representación del estado del arte i-vector[41]. Se mide la eficacia de verificación con HTER[42].

  La figura 9 muestra el comportamiento del HTER para cada sub-banda para a) NB: banda estrecha y b) WB: banda ancha.

a)



b)

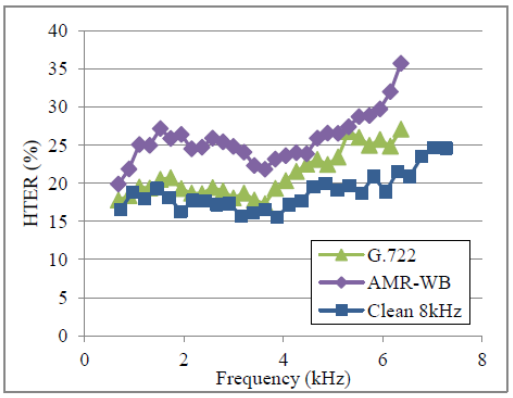


Fig. 9. HTER for subbandas para a) NB: banda estrecha b) WB: banda ancha [16].

El comportamiento superior de la voz sin comprimir sobre la voz codificada, se observa en ambas figuras, así como el mejor comportamiento en general de los códec WB sobre los códec NB, las bandas más baja (50-200Hz) y más alta (3,4-7kHz) que se utilizan en WB y no en NB

---

[39] Regiones de frecuencia con mayor intensidad en el espectrograma de voz, correspondientes a las resonancias del tracto vocal, por lo que caracterizan a cada persona. Para voz masculina por canal telefónico pueden observarse hasta cuatro formantes.

[40] Hyon S., Wang, H., Wei, J., Dang, J., "An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution," Signal and Information Processing Association Annual Summit and Conference, pp.1-4, 2012.

[41] Método de representación vectorial de una muestra de voz para el reconocimiento de locutores, a partir de un modelo UBM, de una matriz con la variabilidad de sesión presente en las muestras de voz de una población y de la estadística de primer y segundo orden de los rasgos de la muestra de voz

[42] Half Total Error Rate: razón de error total medio, medida de eficacia para verificación de locutor obtenida de la curva de comportamiento del error de detección.

contribuyen al mejor comportamiento de la verificación del locutor, lo que puede confirmarse en la tabla 5, que muestra el resultado de los experimentos promediando la eficacia medida en HTER, de todas las sub bandas.

Tabla 5. Eficacia en HTER promedio para todo el espectro [16].

Distortion	HTER (%) TIMIT
G.711 (NB)	6.29
AMR-NB (NB)	8.56
Uncoded 4kHz	3.75
G.722 (WB)	2.03
AMR-WB (WB)	2.89
Uncoded 8kHz	0.88

Si se compara el comportamiento de eficacia para la voz sin comprimir hasta 4 kHz en ambos casos, se observa un mejor comportamiento en el experimento con WB que con NB, al digitalizarse la voz al doble de la frecuencia, lo que se confirma además con el mejor comportamiento de la eficacia de los códec WB sobre los NB en dicha banda de frecuencias. La información contenida en la voz por encima del ancho de banda del canal telefónico (3,4-7kHz) contiene información identificativa del locutor, que se desecha en los códec NB y si se tiene en cuenta en los códec WB.

Con relación a los códec NB (Figura 9a) se observa que el comportamiento del G.711 se acerca bastante al de la voz sin comprimir, mientras que el AMR-NB se aleja algo, esto se debe a que el método de compresión usado por G.711 es menos complejo y el "bit rate" mucho más alto (PCM) que el de AMR-NB, y por consiguiente, se conserva más información del locutor. Algo similar sucede en los códec WB (Figura 9b) donde el comportamiento del G.722 se acerca más al de la voz sin comprimir que el AMR-WB, por la misma razón.

- Los mismos autores en otro trabajo del 2014 [18], evalúan dos tipos de rasgos cepstrales, en escala mel MFCC y en escala lineal LFCC, en verificación de locutores con representación i-vector, utilizando los mismos códec, para voces masculinas y femeninas. Los resultados de eficacia de verificación obtenidos, medidos con HTER, pueden verse en la tabla 6.

Tabla 6. Eficacia en HTER para diferentes condiciones de compresión y diferentes rasgos [18].

Condición	**Masculino**		**Femenino**	
sin códec:	MFCC	LFCC	MFCC	LFCC
0 – 4 kHz	3.63	3.61	8.83	5.06
4 – 8 kHz	5.75	3.82	5.41	5.52
0 – 8 kHz	2.14	1.33	2.18	2.19
con códec:				
G.711 (NB)	7.96	7.00	16.70	11.48
AMR-NB	9.69	9.52	16.98	12.52
G.722 (WB)	2.31	2.27	3.98	3.90
AMR-WB	3.45	4.57	5.63	4.25

Con relación al comportamiento de los rasgos en las diferentes bandas:

- Los rasgos LFCC superan a los MFCC con voces masculinas en todas las condiciones sin códec y usando códec NB y superan apreciablemente a los MFCC con voces femeninas en la

condición sin códec en la banda inferior (0-4 kHz) y usando los códec NB. O sea, los rasgos LFCC presentan mayor eficacia para la voz de ambos sexos, en banda estrecha (0-4 kHz), debido a la distribución lineal de las frecuencias del rasgo LFCC en dicha banda.

- Por otra parte, en las voces femeninas, donde el espectro alcanza valores más altos, los rasgos MFCC superan a los LFCC en la banda superior (4-8 kHz) y consecuentemente en toda la banda (0-8 kHz), debido la distribución mel de las frecuencias del rasgo MFCC en dicha banda.
- Sin embargo, se observa que para los códec BW, los rasgos LFCC brindan una mejor eficacia que los MFCC, en sentido general.

Con respecto al contenido de información identificativa en las diferentes bandas:

- La información identificativa del locutor contenida en la banda superior (4-8 kHz) es similar a la contenida en la banda inferior (0-4 kHz), lo que se observa en la eficacia obtenida para voz sin códec, especialmente para rasgos LFCC. Consecuentemente la eficacia obtenida para la banda completa (0-8 kHz) y para los códec BW es superior a la obtenida para la banda inferior (0-4 kHz) y los códec NB, respectivamente.
- Con relación a la capacidad que poseen los rasgos extraídos para contener dicha información identificativa, se observa que los rasgos LFCC brindan mejor eficacia que los MFCC en la banda inferior (0-4 kHz) y en los códec NB, más acentuado en las voces femeninas. En la banda superior (4-8 kHz) se cumple lo anterior pero solo en las voces masculinas, atribuible a la información identificativa que aparece en voces masculinas alrededor de los 6 kHz.

4.3 Métodos para compensar el efecto de la compresión del códec utilizando la normalización de la puntuación

- En un trabajo publicado en el 2001 [19] los autores investigan el efecto de los códec de banda estrecha GSM EFR (12.2 kb/s), G.729 (8 kb/s), G.723.1 (5.3 kb/s) y MELP (2.4 kb/s)[43], en verificación de locutores, utilizando rasgos MFCC y sus derivadas y un modelo GMM-UBM-MAP de 2048 mezclas, confirmando lo ya observado en [10] y [14]: la pérdida de eficacia del reconocimiento se incrementa a medida que el códec es de menor calidad y es más bajo su "bit-rate", además dicha pérdida se incrementa si la condición de desigualdad del reconocimiento es no pareja.

  En este trabajo se proponen dos métodos de normalización de la puntuación del reconocimiento dependientes del "handset"[44] : HNorm[45] y HTNorm[46] los cuales incrementan la eficacia del reconocimiento y reducen el efecto de la condición de desigualdad no pareja.

  Se proponen cuatro condiciones de desigualdad entre entrenamiento y prueba:
  - "A": completamente pareja, todas las muestras para el modelo UBM, el entrenamiento y la prueba son procesadas por el mismo códec.
  - "B": parcialmente no pareja, las muestras para el modelo UBM y el entrenamiento son procesadas por el mismo códec y las muestras para la prueba no son procesadas. En este caso es una condición no pareja entre entrenamiento y prueba.

---

[43] Mixed Excitation Linear Prediction: códec de muy bajo bit-rate que sintetiza el habla a 2.4 kb/s utilizando armónicos y ruido. El códec MELP es una versión modificada y enriquecida del algoritmo CELP. Es un estándar militar MIL-STD-3005 y se utiliza en radio de banda estrecha, comunicaciones seguras y comunicaciones satelitales.

[44] Se refiere a la tecnología del micrófono del aparato telefónico, en este caso carbón o electrolítico.

[45] Handset Normalization : normalizacion de la puntuación dependiente del handset

[46] Handset Test Normalization: normalización de la puntuación dependiente del handset de la prueba

- "C": parcialmente no pareja, las muestras para el modelo UBM y la prueba no son procesadas y las muestras para el entrenamiento son procesadas por un códec. En este caso es una condición no pareja entre entrenamiento y prueba, similar a la condición "B".
- "D": completamente no pareja, las muestras para el entrenamiento y la prueba son procesadas por el mismo códec pero las muestras para el modelo UBM no son procesadas. En este caso es una condición pareja entre el entrenamiento y la prueba y una condición no pareja entre el modelo UBM, el entrenamiento y la prueba. Es la peor condición de desigualdad.

En la figura 10 se observan el efecto de la no normalización y de los dos métodos de normalización sobre la eficacia de la verificación de locutores medido en EER, para las cuatro condiciones de desigualdad entre entrenamiento y prueba, para los cuatro códec:

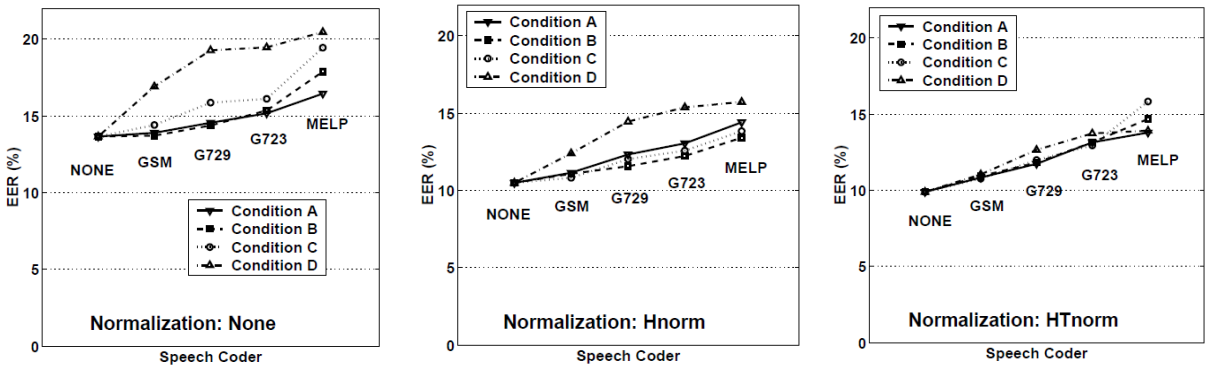


Fig. 10. Efecto de la normalización de la puntuación del reconocimiento para cuatro condiciones de desigualdad [19].

Se observa de la figura:

- En todos los casos se confirma que la eficacia empeora a medida que el "bit-rate" disminuye.
- En todos los casos se observa que las condiciones "B" y "C" de desigualdad brindan eficacias muy similares, peores que la "A" pero mejores que la "D".
- Se observa que el método de normalización HNorm, con respecto al no normalizado, mejora la eficacia para los cuatro códec en las cuatro condiciones de desigualdad, incluso se observa una mejor eficacia de las condiciones "B" y "C" , con respecto a la condición "A", para los cuatro códec.
- Por último se observa que el método de normalización HTNorm, con respecto al no normalizado y al HNorm, mejora aún más la eficacia para tres de los códec (excepto el MELP) en las cuatro condiciones de desigualdad, observándose en dichos códec una eficacia prácticamente igual, independiente de la condición de desigualdad, siendo éste el método de normalización más robusto.

- En un trabajo publicado en 2012 [20] los autores evalúan otro método de normalización de la puntuación del reconocimiento dependiente de la prueba, conocido como TNorm[47] en un experimento de verificación del locutor con rasgos LPCC mas sus derivadas usando un modelo GMM-UBM-MAP con 128 mezclas y utilizando el EER como medida de eficacia.

[47] Test Normalization : Normalización de la prueba

Se comprueba la pérdida de eficacia con respecto a una prueba sin códec (EER: 7%) al comprimir las expresiones de voz utilizando el códec G.729 (8 kb/s) (EER: 17%) y como el método T-Norm logra incrementar de nuevo la eficacia del reconocedor, hasta un EER: 14%.

4.4 Métodos para compensar el efecto de la compresión del códec utilizando la plataforma i-vector de representación de la voz

En un trabajo del 2013 [21] se analiza la influencia de la compresión de diferentes códec en la verificación de locutores, utilizando el método de representación de la voz del estado del arte: i-vector.

Se proponen además varios métodos de mitigación del efecto:

- la utilización de los rasgos MFCC y de dos nuevos tipos de rasgos de la voz, robustos al ruido: MDMC[48] y PNCC[49].
- la clasificación de los locutores con el método PLDA[50] entrenado bajo cuatro condiciones de entrenamiento diferentes: voz limpia (del inglés: "Clean"), voz con ruido y reverberación (del inglés: "Noise & Reverb"), voz codificada en condición pareja con todos los códec (del inglés: "All codec") y voz codificada en condición no pareja, con todos los códec excepto el usado para la prueba (del inglés: "Unseen codec").
- la fusión de los resultados de la clasificación para los tres rasgos utilizados (del inglés: "iv-Fusion").

Se evalúa el efecto de varios códec, algunos utilizados en VoIp como AMR-NB (4.75 a 12.2 kb/s), GSM FR 06.10 (13 kb/s) y Speex (8 kb/s) y otros utilizados para la compresión de audio como AAC (8 y 16 kb/s)[51], MPEG-2 audio layer III (MP3)[52], RealAudio[53] y WMA[54].

En la tabla 7 se observan los resultados de eficacia de verificación promedio entre todos los códec, medida en EER para las cuatro condiciones de entrenamiento del PLDA, para los tres tipos de rasgos usados y su fusión:

Tabla 7. Eficacia en EER promedio de todos los códec para los tres rasgos y las cuatro condiciones de entrenamiento del PLDA [21].

PLDA	MFCC	MDMC	PNCC	ivFusion
Clean	3.06%	2.63%	2.76%	2.37%
Noise & Reverb	2.93%	2.50%	2.61%	2.16%
Unseen Codecs	2.97%	2.63%	2.51%	2.17%
All Codecs	1.81%	2.08%	1.95%	1.56%

Se observa que:

- la eficacia ante la voz comprimida mejora solo marginalmente cuando se entrena el PLDA con voz ruidosa y reverberada ("Noise & Reverb") y con voz en condición no pareja ("Unseen

[48] Medium Duration Modulation Cepstrum: rasgos cepstrum de modulación de media duración
[49] Power Normalized Cepstral Coefficients: coeficientes cepstrales normalizados en potencia
[50] Probabilistic Linear Discriminant Analysis: análisis discriminante lineal probabilístico, método para la compensación de la puntuación en clasificación con la representación i-vector
[51] Advanced Audio Coding: codificación de audio avanzada, método de codificación del audio sustituto del MP3
[52] Moving Picture Experts Group 2 capa de audio III : método de codificación de audio conocido como MP3
[53] Audio Real: Formato de audio propietario de Real Networks
[54] Windows Media Audio: Método de compresión de audio desarrollado por Microsoft

codec"), la mejora apreciable se observa cuando el PLDA se entrena con voz en condición pareja ("All codec").

- el entrenamiento del PLDA con voz en condición no pareja, donde la prueba se hace con un códec no entrenado, no mejora la eficacia con respecto al entrenamiento solo con ruido y reverberación.
- los rasgos robustos a ruido ofrecen una mayor robustez que los MFCC ante la voz comprimida demostrando así su capacidad de generalización ante la compresión de la voz.
- en la fusión de los i-vector de cada rasgo se logra un incremento de la eficacia con respecto a la eficacia para cada rasgo por separado, para las cuatro condiciones de entrenamiento del PLDA.

4.5 Métodos de reconocimiento del locutor utilizando los parámetros contenidos en el flujo de bits del códec

El incremento en el uso del reconocimiento del locutor sobre la VoIP ha conllevado a la aparición de servicios de reconocimiento de locutores sobre Internet. Estos servicios se basan en una arquitectura cliente-servidor y los métodos de reconocimiento residen en la red. El cliente de este servicio, que desea autenticarse por voz, la envía a través de la red, la misma se comprime con un códec y llega hasta el servidor donde se lleva a cabo el reconocimiento del locutor utilizando los parámetros contenidos en el flujo de bits del códec [22] (ver figura 7 pto. 4), este método se conoce como reconocimiento del locutor en dominio comprimido (del inglés "Compressed Domain") [23].
El reconocimiento convencional del locutor no puede aplicarse a la VoIP de forma inmediata, ya que se requiere re-sintetizar la voz a la salida del códec en el extremo receptor (ver figura 7 pto. 2), para obtener una voz similar a la trasmitida, convertirla a un formato PCM y extraerle los rasgos, para solo entonces reconocer al locutor propietario de la voz.
En este nuevo servicio, se utiliza para hacer el reconocimiento del locutor en dominio comprimido los parámetros contenidos en el flujo de bits de la voz comprimida al llegar al extremo receptor (el servidor), en lugar de extraer los rasgos de la voz re-sintetizada, evitándose un proceso de síntesis y análisis de la voz innecesario, logrando una reducción de la demora en la respuesta del sistema de reconocimiento, lo que facilita el procesamiento de muchos clientes de forma simultanea [8].
En un futuro ya cercano, con el incremento de la telefonía sobre VoIP, se incrementará el interés en el reconocimiento del locutor en dominio comprimido, para brindar diversos servicios autenticados por voz y mayores facilidades a aquellas entidades gubernamentales que trabajan directamente monitoreando los flujos de datos de la red y que requieran identificar personas por la voz en VoIP [12].

4.5.1 Reconocimiento del locutor utilizando el flujo de bits del algoritmo CELP

El algoritmo CELP[55] es utilizado en el códec G.729, y brinda una calidad adecuada teniendo en cuenta su reducido uso del ancho de banda del canal (bit rate: 6.4, 8 y 11.8 kb/s), siendo ampliamente utilizado en VoIP. Este algoritmo extrae los parámetros LPC de la voz en tramas de 10 ms. en forma de líneas espectrales (del inglés "LSF: line spectrum frequency"), las que se ha comprobado son más robustas al ruido del canal.

- En trabajos publicados en el 2000 [24], [25] los autores evalúan la eficacia de los parámetros del códec G.729 en reconocimiento de locutores en dominio comprimido, extrayendo rasgos cepstrales

[55] Code-Excited Linear Prediction, método de codificación propuesto por M. R. Schroeder y B. S. Atal en "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," en *Proceedings of the IEEE ICASSP 1985*. También es conocido como LPC-10.

en escala Mel a partir de los parámetros LSF y comparando con los rasgos MFCC extraídos de la señal de voz re-sintetizada. Se obtienen diversas variantes de rasgos cepstrales, agregando el pitch y la energía, normalizados o no, pero no se logra en ningún caso superar la eficacia de reconocimiento de locutores utilizando los MFCC de la voz re-sintetizada.

- En el 2003 [26] los autores evalúan la eficacia para verificación de locutores en dominio comprimido, de tres tipos de parámetros obtenidos de los flujos de datos de códec G.729 y G.723.1[56]
  a) rasgos LFCC obtenidos de los parámetros LPC en el codificador del códec
  b) rasgos LFCC obtenidos de los parámetros LSF en el decodificador del códec
  c) rasgos LFCC obtenidos de los parámetros LSF en el decodificador del códec + los rasgos LPCC obtenidos del residuo de los parámetros LPC ante condiciones de desigualdad entre entrenamiento y prueba.

  Se pudo comprobar que, principalmente en condición de desigualdad "no pareja", el rasgo c) es el que mejor eficacia promedio brinda, y en segundo lugar el rasgo b), lo que confirma que:
  - la información residual de los parámetros LPC contiene información identificativa de las personas, que normalmente se descarta.
  - la transformación que lleva a cabo el códec, de los parámetros LPC a LSF, eleva la eficacia del reconocimiento de locutores.

  Este trabajo no compara sus resultados con la eficacia obtenida usando los rasgos de la voz re-sintetizada.
- En el 2005 [27] los autores proponen un método de representación de la voz codificada con el códec G.729 que denominan "CSR" (del inglés: "Compressed Speaker Recognition"). El método CSR forma nuevos vectores de rasgos que contienen información de cada locutor, a partir de los parámetros LSF, el pitch y la energía de cada trama de voz codificada y obtiene una representación de histograma para cada locutor utilizando técnicas de agrupamiento.

  Este método es muy rápido, lo que permite realizar el reconocimiento de varios locutores simultáneamente sobre el flujo de datos de la red, seleccionando los paquetes de VoIP, obteniendo las representaciones, agrupándolas y clasificando por comparación de histogramas. Los autores reportan una eficacia de reconocimiento del 80% y una velocidad tres veces superior a un método utilizando clasificadores GMM.
- Ese mismo año otros autores [28] realizan un estudio de la eficacia y robustez que se obtiene en verificación de locutores en dominio comprimido, con diferentes combinaciones de parámetros del códec G.729, para diferentes condiciones de desigualdad entre entrenamiento y prueba, debido a la adición de ruido blanco a la voz.

  Se evaluaron cinco combinaciones de parámetros del códec, dos obtenidas de la voz sin codificar o re-sintetizada (Set A) y tres obtenidas de la voz codificada (Set B), en todos los casos el ruido fue incorporado a la señal de voz a la entrada del códec con niveles de SNR[57] de 10 dB, 15 dB, 20 dB y sin ruido (del inglés: "Clean"). Las cinco combinaciones de parámetros del códec fueron:
  - A.1: experimento línea base, rasgos MFCC con sus derivadas, sin codificar.
  - A.2: iguales rasgos a los de A.1 pero obtenidos de la voz re-sintetizada después de haber sido procesada por el códec.
  - B.1: rasgos LPC con sus derivadas, obtenidos a partir de los parámetros LSF de la voz codificada.
  - B.2: rasgos MFCC con sus derivadas, obtenidos de los parámetros LSF de la voz codificada.

[56] Utiliza una variante del algoritmo CELP conocida como Algebraic-CELP a más bajo bit rate: 5.3 y 6.3 kbit/s

[57] Signal to Noise Ratio: relación señal-ruido, es una medida de la relación entre la potencia de señal y la potencia de ruido presentes en una señal de voz.

- B.3: iguales rasgos a los de B.2 pero incorporando la información del pitch obtenida del residuo de los parámetros LPC.

Para todas las combinaciones de rasgos se utilizó el modelo GMM con 64 mezclas. En la tabla 8 se observan los resultados de % de EER para el experimento de verificación, realizado bajo condición de desigualdad pareja de ruido para entrenamiento y prueba, columnas 20, 15, 10, y bajo condición de desigualdad no pareja de ruido, con entrenamiento "cln" y prueba con ruido, columnas (20), (15), (10).

Tabla 8. Eficacia de la verificación de locutores en EER para diferentes tipos de rasgos y condiciones de ruido.

Exp /dB	Clean	20	15	10	(20)	(15)	(10)
A.1	0.18	1.0	1.4	2.4	3.8	12.0	22.0
A.2	0.41	1.5	2.2	3.1	3.1	11.0	21.0
B.1	0.34	-	1.6	2.8	15.0	28.0	-
B.2	0.45	1.2	1.9	3.3	3.8	10.0	21.0
B.3	0.34	1.0	1.5	2.5	2.7	7.2	17.0

Los resultados obtenidos sugieren que:

- Aunque el ruido deteriora la eficacia de la verificación, el efecto del ruido enmascara el efecto del códec, al observarse % de EER similares entre el set A y el set B bajo condición de desigualdad pareja de ruido.
- Los modelos de locutores entrenados con voz re-sintetizada (A.2) son más robustos que los modelos entrenados con voz sin codificar (A.1) y con voz codificada (B.1 y B.2), bajo condición de desigualdad no pareja de ruido.
- Para parámetros obtenidos del códec (B.1 y B.2), los LPC se comportan mejor cuando la voz no tiene ruidos y en condición de desigualdad pareja, sin embargo en condición de desigualdad no pareja, los rasgos MFCC superan a los LPC, siendo más robustos a condiciones de desigualdad.
- La incorporación del pitch, obtenido del residuo de los parámetros LPC (B.3), confirma que el mismo contiene información discriminativa del locutor, que no se tiene en cuenta en las demás combinaciones de codificación, y que se refleja en una mejor eficacia de verificación de locutores utilizando ese parámetros, para ambas condiciones de desigualdad. Incluso si se compara con la línea base (A.1) sin codificar presenta mejor eficacia en condición no pareja de ruido y similar eficacia en condición pareja de ruido.
- En sentido general, las combinaciones obtenidas de la voz codificada (B.1, B.2, B.3), brindan mejores resultados que la obtenida con la voz re-sintetizada (A.2).

A partir de los resultados de este trabajo puede asegurarse que la eficacia del reconocimiento de locutores en dominio comprimido, o sea, utilizando parámetros de la voz codificada con el códec G.729, es similar a la obtenida utilizando rasgos de la voz re-sintetizada, incluso brinda una robustez al ruido similar. Para lograr lo anterior deben seleccionarse adecuadamente los parámetros y llevar a cabo su conversión a rasgos MFCC.

4.5.2 Reconocimiento del locutor utilizando el flujo de bits del códec GSM

El ETSI[58] ha propuesto tres algoritmos de codificación para la telefonía móvil GSM [5], todos soportados sobre diferentes variantes de los parámetros LPC de la voz en tramas de 20 ms.:

- GSM FR 06.10 (13 kb/s) que utiliza parámetros RPE-LTP[59] LPC, con una calidad pobre de la voz, pero en su momento (año 1987) fue la mejor opción de codificación GSM.

[58] European Telecommunications Standards Institute: Instituto Europeo de Estándares de Telecomunicaciones

- GSM HR 06.20 (5.6 kb/s) que utiliza parámetros VSELP[60] que es una variante de menor ancho de banda con mayor costo de procesamiento pero aun con baja calidad de voz.
- GSM EFR 06.60 (12.2 kb/s) que es una variante mejorada del GSM FR pero basado en el algoritmo Algebraic CELP, que brinda mejor calidad de voz y es más robusto a las dificultades que se presentan en la red.

- En dos trabajos publicados en el 2000 [15, 29] se evalúa la eficacia de reconocimiento de locutores en dominio comprimido para los parámetros que brindan los códec GSM FR y GSM EFR, convertidos a rasgos LPCC y evaluados con un clasificador GMM, en identificación y verificación de locutores, comprobándose que:
  - El uso de rasgos LPCC obtenidos de los parámetros LPC que entrega el códec, no afecta la eficacia del reconocimiento, comparado con el uso de los rasgos LPCC obtenidos de la señal de voz re-sintetizada.
  - La incorporación de la energía, obtenida del residuo de los parámetros LPC, es vital para obtener dicha eficacia.
  - El orden del filtro que brinda los parámetros LPC afecta dicha eficacia.

- En trabajos posteriores del 2005 [30] y 2006 [12], los autores evalúan la eficacia de nueve parámetros que brinda el algoritmo Algebraic CELP utilizado en el códec GSM AMR[61] (12.2 kb/s), para cuatro duraciones diferentes de trama del flujo de bits (5, 10, 15 y 20 s.), utilizando como criterio de identificación la fusión de la distancia euclideana cuadrada entre dos de las estadísticas (covarianza y skewness) de la trama, en un sencillo experimento de identificación de locutores en dominio comprimido, con 14 locutores.En dichos trabajos se analiza la capacidad identificativa de diferentes parámetros de dicho códec, seleccionando los más identificativos con F-Ratio[62]. Se evalúan 18 parámetros agrupados en tres variantes REF-18: todos, TOP-9: los mejores 9 según FRatio y BEST-9: los nueve mejores obtenidos de N comparaciones con N-1 parámetros. La tabla 9 muestra los resultados.

Tabla 9. Eficacia de identificación para diferentes duraciones del flujo de bits de prueba y las tres combinaciones de rasgos [12].

Length (s)	REF-18	TOP-9	BEST-9
5	71.74%	80.54%	83.68%
10	87.72%	92.12%	92.72%
15	96.05%	95.39%	98.16%
20	100%	96.25%	100%

Se comprueba que:
- A medida que se incrementa la duración de la trama de bits a la que se le analiza su estadística, se logra alcanzar la máxima eficacia de identificación, al llegar a 20 s.

---

[59] Regular Pulse Excitation –Long Term Prediction: variante de predicción lineal a largo término.
[60] Vector Sum Excited Linear Prediction: variante del método CELP que realiza la predicción lineal de forma vectorial.
[61] Conocido también como AMR-NB: es una variante adaptiva multi-rate del algoritmo Algebraic-CELP, de bajo bit rate y con alta calidad de la voz.
[62] Razón de varianzas. El F-ratio de un rasgo se computa como la razón entre su varianza entre-locutores y el promedio de su varianza intra-locutor.

- De los 18 parámetros que entrega el algoritmo Algebraic-CELP, solo son necesarios nueve para lograr similar eficacia de identificación, reduciendo a la mitad la memoria necesaria para almacenar los parámetros.
- La medida FRatio no da una idea exacta de la capacidad identificativa de los parámetros, cuando se evalúan de conjunto. Para cualquier duración, se logra una mejor eficacia para BEST-9 que para TOP-9.
- Este método es muy eficiente y posee poco costo computacional, asegurando con 20 s. de voz codificada identificar al locutor "en vivo".

Se requiere por supuesto incrementar el volumen de locutores a identificar para una mejor evaluación del método.

4.5.3 Reconocimiento del locutor utilizando flujos de bits de otros códec

El incremento de la telefonía sobre VoIP con el consiguiente incremento en el interés por evaluar el reconocimiento del locutor "en vivo", en dominio comprimido, para diferentes algoritmos de compresión, ha conllevado a la realización de investigaciones sobre otros códec.

- En otro trabajo publicado en el 2006 [23], que es una generalización del [12], los autores evalúan la eficacia en reconocimiento del locutor de los parámetros de cuatro algoritmos asociados a diez códec diferentes utilizados en VoIP, con diferentes técnicas de compresión y muy bajos bit-rates:
  - basados en el algoritmo Algebraic CELP, utilizado en el códec GSM AMR (4.75, 6.7, 7.4 y 12.2 kb/s), G.729 (8 kb/s) y G.723.1 (5.3 kb/s)
  - basados en LPC, utilizado en el códec iLBC (13.3 kb/s y 15.2 kb/s)
  - basados en MPC-MLQ[63] utilizado en el códec G.723.1 (6.3 kb/s)
  - basados en algoritmo MELP (2.4 kb/s) utilizado en el códec del mismo nombre.

  Los experimentos de reconocimiento de locutor utilizando combinaciones de parámetros de dichos códec, se llevaron a cabo con tres duraciones diferentes de trama del flujo de bits (10, 20 y 30 s.), utilizando el mismo criterio de identificación de [12] y [30], en un experimento de identificación de locutores en dominio comprimido, con 14 locutores. La tabla 10 refleja los resultados de la identificación, donde N es la cantidad de parámetros extraídos del flujo de bits, utilizados para clasificar.

Tabla 10. Eficacia de identificación para diferentes duraciones del flujo de bits de prueba [23].

Speech Coder	Length (s)			N
	10	20	30	
GSM AMR 12.2	95.8%	100%	100%	9
GSM AMR 7.40	85.2%	91.9%	95.3%	6
GSM AMR 6.70	85.5%	93.8%	97.1%	6
GSM AMR 4.75	84.9%	92.5%	96.2%	6
G.729 8.00 kb/s	77.0%	83.1%	87.7%	7
G.723.1 6.3 kb/s	76.4%	85.0%	90.6%	6
G.723.1 5.3 kb/s	75.5%	86.1%	87.6%	6
MELP 2.4 kb/s	86.1%	95.6%	97.2%	9
iLBC 15.20 kb/s	75.8%	82.5%	92.5%	10
iLBC 13.33 kb/s	77.9%	89.4%	95.3%	13

[63] Multipulse LPC with Maximum Likelihood Quantization: variante de predicción lineal con cuantificación vectorial

De este último estudio se puede concluir:

- se valida la efectividad del reconocimiento del locutor en dominio comprimido para diferentes algoritmos, "bit-rate" y parámetros de compresión utilizados, aunque la data de prueba es muy pequeña.
- Los algoritmos utilizados en GSM AMR, para diferentes "bit rate", alcanzan los más altos valores de identificación (> 90%) en menores tiempos (20 s.).
- Se observa para los códec GSM AMR una disminución en la eficacia a medida que se reduce N, el número de parámetros escogidos para comparar. Es de esperar que la eficacia puede incrementarse tomando más parámetros del flujo de bits, cambiando la medida de distancia o incrementando la duración de la trama de bits a procesar.
- Se observa que la eficacia de identificación no depende del "bit-rate", el códec MELP (2.4 kb/s) y el códec GSM AMR (4.75 kb/s), que son los de más bajo "bit rate", tienen una eficacia superior a otros que los superan en "bit-rate" y en número de parámetros como los códec iLBC (13.33 y 15.20 kb/s), G.723.1(5.3 y 6.3 kb/s) y G.729 (8 kb/s). Evidente estos dos algoritmos extraen más información identificativa del locutor en sus parámetros, en menos parámetros y a menos "bit-rate".
- En sentido general para todos los códec evaluados, la eficacia de identificación de este método de reconocimiento de locutores en dominio comprimido depende del número de parámetros y de la duración de la trama del flujo de bits utilizados así como del criterio de distancia, asociado éste a la estadística de las tramas. Se observa que en general para todos los códec evaluados se logra alcanzar una alta eficacia en la identificación con 30 s. de voz de la persona.

5 La VoIP utilizada para el reconocimiento forense del locutor

Aunque este reporte está dedicado a evaluar como la QoS de la VoIP influye en el reconocimiento automático del locutor, creemos necesario hacer un breve análisis de su influencia en el reconocimiento forense del locutor, aspecto que ha comenzado a estudiarse también.

En el campo forense, el reconocimiento del locutor es una herramienta de gran valor identificativo, cuando sólo se cuenta con las voces de los implicados en un delito. La integración de la VoIP con la telefonía PSTN ha creado nuevos retos nunca antes vistos en el campo del cumplimiento de la ley, al extenderse su utilización por los delincuentes en delitos como: homicidios, drogas, secuestros, amenazas de bombas, asaltos, llamadas obscenas o falsas, crímenes de collar blanco, fraudes bancarios, etc. [2].

5.1 Influencia de la compresión de los códec en las características espectrales de la voz

En un estudio llevado a cabo recientemente y publicado en el 2013 [31] se comprueba que el proceso de compresión-descompresión que sufre la voz debido a su paso por los códec, provoca distorsiones y compresiones en las características espectrales de la voz, que se aprecian espectralmente y que además se pueden medir en el comportamiento de los formantes en el espectrograma.

Se evalúan los dos tipos de interfaz[64] más comunes de acceso de la voz a Internet: telefonía PSTN integrada a TCP/IP ("PSTN") y VoIP entre computadoras ("PC") en sus diferentes combinaciones PSTN-PSTN, PC-PC y PC-PSTN, y que son comparadas con el interfaz "IP-phone", llegando a las siguientes conclusiones:

- En PSTN-PSTN se pierden formantes y otros se descontinúan.

[64] Según la agrupación realizada al finalizar el epígrafe 1.

- En PC-PSTN, el segundo formante se desplaza y se une al primer formante.
- En PSTN-PSTN se observan que los formantes tres y cuatro se “comprimen” en frecuencia.
- En PC-PSTN se observa que los cuatro formantes se “comprimen” ligeramente en frecuencia.

Se puede observar en el estudio, que el interfaz obtenido del proceso de integración de la telefonía PSTN a la plataforma VoIP es el que más afecta la calidad y la distribución espectral de los formantes. Aunque en el trabajo no se especifican los códec utilizados en cada interfaz, estos son distintos, de ahí el diferente comportamiento de los formantes. Esta información puede ser de gran valor para el reconocimiento forense de locutores, donde el perito compara el comportamiento de los formantes de dos muestras de voz para determinar su similitud y establecer la identidad de un locutor desconocido.

5.2 Influencia de diferentes parámetros del QoS de la VoIP sobre la identificación auditiva del locutor por oyentes familiarizados

En un trabajo también reciente, publicado en el 2013 [32], los autores evalúan como el uso de diferentes interfaces electro-acústicos, anchos de banda y esquemas de codificación de los códec y diferentes pérdidas de paquetes, influyen en la identificación auditiva de locutores por oyentes familiarizados, muy similar al método forense conocido como “voice line-up”[65].

Se evaluaron cuatro interfaces de usuarios, representativos de diferentes servicios de VoIP, asociados a diferentes códec y perdidas de paquetes, en transmisión y en recepción: teléfono (“phone with handset”), teléfono manos libres (“hands-free phone”), auricular con micrófono (“headset”) y teléfono móvil (“mobile phone”).

Se evalúan los códec de banda estrecha “NB” G.711 (64 kb/s) y AMR-NB (12.2 kb/s), de banda ancha “WB” G.722 (64 kb/s) y AMR-WB (12.65 kb/s) y de banda super-ancha, “SWB”[66] G.722.1 (32 y 48 kb/s). Solo se evalúa la pérdida de paquetes en las comunicaciones que utilizan el teléfono. En la tabla 11 se muestran las diferentes combinaciones evaluadas.

Se grabaron expresiones leídas de voces de 16 locutores (8 hombres y 8 mujeres) que fueron después reproducidas a través de los diferentes interfaces de usuarios usando un simulador profesional de cabeza y torso, y posteriormente trasmitidas a través de un simulador VoIP, donde se programaron los diferentes códec y las pérdidas de paquetes. Un grupo de 20 personas (16 hombres y 4 mujeres), colegas de los anteriores (familiarizados con las voces) integraron el tribunal de audición, 10 de ellas (6 hombres y 4 mujeres) fueron locutores, que tuvieron que identificar sus propias voces procesadas.

[65] Alineación de voces, método forense donde los peritos comparan auditivamente varias voces

[66] Ver epigrafe 3.3.2

Tabla 11. Interfaces de usuarios evaluados, con sus correspondientes códec y perdida de paquetes. [32].

Interface	Codec	bit rate (kbps)	Packet loss
Phone with handset (SNOM 870) 	G.711 (A-law) (NB)	64	0, 5, 10, 15
	G.722 (WB)	64	0, 5, 10, 15
Hands-free phone (Polycom IP 7000) 	G.711 (A-law) (NB)	64	0
	G.722 (WB)	64	
Headset (Beyerdynamic DT 790) 	G.711 (A-law) (NB)	64	0
	G.722 (WB)	64	
	G.722.1C (SWB)	32	
	G.722.1C (SWB)	48	
Mobile phone (SONY XPERIA T) 	AMR-NB (NB)	12.2	0
	AMR-WB (WB)	12.65	

Se comprobó que al pasar de canal NB a canal WB se incrementa la exactitud en la identificación cuando se trasmite por cualquiera de los cuatro interfaces, sobre todo a través del auricular y el teléfono móvil; el canal WB brinda mejor eficacia de reconocimiento cuándo se recibe a través del teléfono manos libres o del auricular y menor cuando se recibe por el teléfono. Sin embargo el paso a SWB en el auricular no provocó una mejora con respecto al WB.

Con relación a la pérdida de paquetes el canal WB comienza a perder eficacia en la identificación a partir del 15% de pérdida mientras que el canal NB ya presenta afectaciones desde el 5% de pérdida.

El estudio confirma que la identificación auditiva por oyentes familiarizados (que pudiera extenderse al método forense "voice line-up") se favorece con un mayor ancho de banda y mayor "bit-rate" de los códec y que el auricular y el teléfono manos libres aprovechan mejor el ancho de banda de la señal facilitando la identificación.

En un trabajo previo [33] de los mismos autores se llevó a cabo un experimento similar sin utilizar los interfaces y con menos códec, en el mismo ya se observa la superioridad de WB sobre el NB en la identificación auditiva de locutores.

En los trabajos ya referidos [16], [17] y [18] los mismos autores extienden este estudio al reconocimiento automático de locutores, donde se confirma lo relacionado con la influencia del ancho de banda y el bit-rate de los códec en la eficacia del reconocimiento.

6 Conclusiones

Como conclusión se intenta resumir cómo la QoS de la VoIP afecta al reconocimiento del locutor, así como cuales han sido algunos de los métodos utilizados para compensar dichas afectaciones.

Con relación a la influencia de la pérdida de paquetes:

- El reconocimiento de locutor independiente del texto es poco sensible a la perdida de paquetes, pudiendo soportar altos niveles relativos de perdidas, sin afectar su eficacia. A partir de que el

modelo GMM considera cada trama de rasgos de forma independiente, no es sensible a la ruptura temporal debido a la perdida de paquetes. Se comprueba que la alta redundancia contenida en las tramas, permite alto % de pérdida de paquetes.

- El reconocimiento de locutor dependiente del texto es más sensible a la perdida de paquetes que el independiente del texto, debido a que la voz en el primero mantiene una relación secuencial-temporal con el texto que se dice y la perdida de dicha relación asociada al texto, provoca perdida de información para el reconocimiento.
- Si la codificación utilizada es de muy bajo "bit-rate", entonces **si** se incrementa la sensibilidad de la verificación de locutor ante la pérdida de paquetes, pudiendo disminuir su eficacia considerablemente.

Con relación a la influencia del códec:

- La eficacia del reconocimiento de locutores empeora cuando la voz se codifica. En sentido general, a mayor nivel de compresión de la voz para un mejor aprovechamiento del ancho de banda del canal de comunicación debido a la reducción del "bit rate", menor eficacia en el reconocimiento.
- La condición pareja de desigualdad entre entrenamiento y prueba brinda la mejor eficacia del reconocimiento de locutores con respecto a cualquier condición no pareja, pero solo para códec con alto "bit-rate". En condición no pareja, llevar a cabo el entrenamiento con una voz codificada con un códec de alto "bit rate" brinda la mejor eficacia en el reconocimiento.
- La eficacia del reconocimiento de locutores depende directamente del ancho de banda de los códec utilizados para comprimir la voz. La banda superior (4-8 kHz) contiene tanta información identificativa del locutor como la banda inferior (0-4 kHz).
- En general los rasgos LFCC son más adecuados que los rasgos MFCC para el reconocimiento del locutor sobre VoIP.
- Los métodos de normalización de la puntuación del reconocimiento elevan en sentido general la eficacia del reconocimiento y reducen el efecto de la condición no pareja de desigualdad entre entrenamiento y prueba.
- Otros métodos propuestos para compensar el efecto de la compresión del códec, que elevan la eficacia del reconocimiento de locutores, son:
    - La utilización de nuevos rasgos robustos ante ruido.
    - La fusión de clasificadores i-vector utilizando diferentes rasgos.
- Los métodos de reconocimiento de locutores en el "dominio comprimido" pueden llegar a ser similares en eficacia y robustez al ruido a los métodos que utilizan la voz re-sintetizada, siempre que se escojan adecuadamente los parámetros del flujo de bits, la duración de la trama de bits y el método de comparación. Se ha comprobado que algunos códec de bajo "bit-rate", de reciente aparición, tienen mejor comportamiento que los de alto "bit-rate", debido a que los parámetros obtenidos por sus algoritmos poseen mayor capacidad identificativa de la persona. Dichos resultados hacen competitivos estos métodos para servicios de autentificación e identificación de locutores en Internet, siempre que se logre una adecuada eficiencia en la explotación.

La introducción de la tecnología de VoIP en las comunicaciones de nuestro país, constituye un nuevo reto para el procesamiento de la voz, específicamente el reconocimiento del locutor, tanto desde el punto de vista forense como de forma automática.

En el CENATAV se han desarrollado diversos métodos para el reconocimiento automático de locutores independiente del texto, los que se han probado con voz por canal telefónico con resultados de eficacia muy buenos. Este estudio nos ha permitido ubicarnos en el tema y precisar hacia donde deben dirigirse los esfuerzos para elevar la eficacia y eficiencia del reconocimiento automático del locutor sobre VoIP.

Como primer paso, se requiere evaluar el comportamiento de nuestros métodos, en proceso de implementación e implantación en aplicaciones y productos, ante la VoIP. Para llevar a cabo dicha tarea

se requiere crear una base de voces VoIP simulando muestras de voz procesadas por una red Internet, utilizando herramientas como las referidas en [34]. Nos encontramos en estos momentos evaluando la factibilidad de ejecución de la misma.

Referencias bibliográficas

1. Singh H.P., et al.: VoIP: State of art for global connectivity: A critical review. Journal of Network and Computer Applications, Vol. 37. (2014) 365–379.
2. Patil H.A. et al.: Chapter 6: Speaker Identification over Narrowband VoIP Networks. In: A. Neustein, H. A. Patil (Eds.): Forensic Speaker Recognition Law Enforcement and Counter-Terrorism. (2012) 125-152.
3. Karapantazis S., Pavlidou F. T.: VoIP: A comprehensive survey on a promising technology. Computer Networks, 53. (2009) 2050–2090.
4. Skoglund J. et al.: Chapter 15: Voice over IP: Speech Transmision over Packet Networks. In: J. Benesty, M. M. Sondhi, Y. Huang (Eds.): Springer Handbook of Speech Processing. (2008) 307-330.
5. Kazemitabar, H. et al.: A Survey on Voice over IP over Wireless LANs. World Academy of Science, Engineering and Technology, 47. (2010) 352-358.
6. Staroniewicz, P., Majewski W.: Methodology of Speaker recognition Test in Semi-real VoIP Conditions. In: Proceedings of Third COST 275 Workshop “Biometrics on the Internet”. (2005) 33-36.
7. Besacier, L. et al.: Voice Biometrics over the Internet in the Framework of COST Action 275. EURASIP Journal on Applied Signal Processing. Vol. 4 (2004) 466–479.
8. Besacier L et al.: Speech coding and packet loss effects on speech and speaker recognition. In: Tan ZH, Lindberg B (Eds.): Automatic speech recognition on mobile devices and over communication networks. Springer, (2008) 27-39.
9. Yajnik, M. et al.: Measurement and modeling of the temporal dependence in packet loss. In: Proceedings of 18th Annual Joint Conference of the IEEE Computer and Communications Societies, Infocom’99. (1999) 345–352.
10. Silovsky, J. et al.: Assessment of Speaker Recognition on Lossy Codecs Used for Transmission of Speech. In: 53rd International Symposium ELMAR-2011. (2011) 205-208.
11. Borah, D.K., DeLeon P.: Speaker Identification in the presence of Packet Losses. In: proceedings of Digital Signal Processing Workshop and the 3rd IEEE Signal Processing Education Workshop. (2004) 302-306.
12. Petracca, M. et al.: Optimal Selection of Bit Stream Features for Compressed-Domain Automatic Speaker recognition. In Proceedings of 17th European of Signal Processing Conference, EUSIPCO. (2006).
13. Janicki A., Staroszczyk T.: Speaker Recognition from Coded Speech Using Support Vector Machines. In: Proceedings of the 14th international Conference on Text, Speech and Dialogue, LNAI 6836. (2011) 291-298.
14. Janicki A.: SVM-based Speaker Verification for Codec and Un-coded Speech. In: Proceedings of 20th. European Signal Processing Conference, EUSIPCO. (2012) 26-30.
15. Besacier L. et al.: GSM Speech Coding and Speaker Recognition. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00. Vol. 2, (2000). 1085 – 1088.
16. Fernández Gallardo L. et al.: Spectral Sub-band Analysis of Speaker Verification Employing Narrowband and Wideband Speech. In: Proceedings of the Speaker and Language Recognition Workshop, Odyssey’14. (2014).

17. Fernández Gallardo L. et al.: Analysis of Automatic Speaker Verification Performance over Different Narrowband and Wideband Telephone Channels. In: Proceedings of SST (2012) 157-160.
18. Fernández Gallardo L. et al.: Advantages of Wideband over Narrowband Channels for Speaker Verification Employing MFCCs and LFCCs. In: Proceedings of International Conference on Speech Technologies, Interspeech'14 (2014).
19. Dunn, R.B. et al.: Speaker recognition from coded speech in matched and mismatched conditions. In Proceedings of the Speaker and Language Recognition Workshop, Odyssey'01. (2001), 72–83.
20. Yessad D., Amrouche A.: Robust regression fusion of GMM-UBM and GMM-SVM normalized scores using G729 bit-stream for speaker recognition over IP. International Journal of Speech Technologies. 17, (2014) 43–51.
21. McLaren M. et al.: Improving Robustness to Compressed Speech in Speaker Recognition. In: Proceedings of International Conference on Speech Technologies, Interspeech'13. (2013) 3698-3701.
22. Yessad, D. et al.: Influence of G729 Speech Coding on Automatic Speaker Recognition in VoIP Applications. In: James Jong Hyuk Park et al. (Eds.): Computer Science and Convergence, LNEE 114, Springer, (2012) 745-751.
23. Petracca, M. et al.: Performance Analysis of Compressed-Domain Automatic Speaker Recognition as a function of Speech Coding Technique and Bit-Rate. In Proceedings of International Conference on Multimedia and Expo, ICME'06. (2006).
24. Quatieri T.F. et al.: Speaker and Language Recognition using Speech Codec Parameters. In: Proceedings of European Conference on Speech Technologies, Eurospeech'99. Vol. 2, (1999) 787-790.
25. Quatieri, T. F. et al.: Speaker Recognition using G. 729 Speech Codec Parameters. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00. Vol. 2, (2000) 1089–1092.
26. W. M. Yu E. et al.: Speaker Verification Based on G.729 and G.723.1 Coder Parameters and Handset Mismatch Compensation. In: Proceedings of European Conference on Speech Technologies, Eurospeech'03. (2003) 1681–1684.
27. Aggarwal C. et al.: CSR: Speaker Recognition from Compressed VoIP Packet Stream. In: Proceedings of IEEE International Conference on Multimedia and Expo, ICME'05. (2005) 970-973.
28. Moreno-Daniel, A. et al.: Robustness of Bit-stream based Features for Speaker Verification. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'05. (2005) 749–752.
29. Grassi S. et al.: Speaker Recognition on Compressed Speech. In: Proceedings of the International COST 254 Workshop on Friendly Exchanging through the Net. (2000) 117-222.
30. Petracca, M. et al.: Low-complexity Automatic Speaker Recognition in the Compressed GSM-AMR Domain. In: Proceedings of IEEE International Conference on Multimedia and Expo ICME'05. (2005) 662–665.
31. Sharma S. et al: Spectrographic Study of Speech Samples Recorded through Voice over Internet Protocol (VoIP). In Proceedings of International Conference Oriental COCOSDA and Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE'13. (2013)
32. Fernández Gallardo L. et al.: Human Speaker Identification of known Voices Transmitted through different User Interfaces and Transmission Channels. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'2013. (2013)
33. Fernández Gallardo L. et al.: Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems. In: proceedings of ITG-Fachbericht 236: Sprachkommunikation. (2012).
34. Chowdhury, P. et al.: A framework for VoIP speech data generation using Asterisk. In: proceedings of International Conference on Devices and Communications, ICDeCom'11. (2011).

Anexo 1. Principales características técnicas de los códec estándar utilizados [3]

Table 3
Characteristics of the most well-known voice codecs.

Codec	Bitrate (kb/s)	Frame (ms)	Bits per frame	Algorithmic delay[a] (ms)	Codec delay[b] (ms)	Compression type	Complexity (MIPS)[c]	MOS
Narrowband codecs								
G.711	64	0.125	8	0.125	0.25	PCM	$\ll 1$	4.1[d]
G.723.1	6.3	30	189	37.5	67.5	MP-MLQ	$\leqslant 18$	3.8
G.723.1	5.3	30	159	37.5	67.5	ACELP	$\leqslant 18$	3.6
G.726	16	0.125	2	0.125	0.25	ADPCM	≈ 1	–
G.726	24	0.125	3	0.125	0.25	ADPCM	≈ 1	3.5
G.726	32	0.125	4	0.125	0.25	ADPCM	≈ 1	4.1
G.728	16	0.625	10	0.625	1.25	LD-CELP	≈ 30	3.61
G.729	8	10	80	15	25	CS-ACELP	$\leqslant 20$	3.92
G.729A	8	10	80	15	25	CS-ACELP	$\leqslant 11$	3.7
G.729D	6.4	10	64	15	25	CS-ACELP	<20	3.8
G.729E	11.8	10	118	15	25	CS-ACELP LPC	<30	4
GSM-FR	13	20	260	20	40	RPE-LTP	≈ 4.5	3.6
GSM-HR	5.6	20	112	24.4	44.4	VSELP	≈ 30	3.5
GSM-EFR	12.2	20	244	20	40	ACELP	≈ 20	4.1
AMR-NB	4.75–12.2	20	95–244	25	45	ACELP	15–20	3.5–4.1
iLBC	13.33	30	400	40	60	LPC	18	3.8
iLBC	15.2	20	304	25	40	LPC	15	3.9
Speex (NB)	2.15–24.6	20	43–492	30	50	CELP	8–25	2.8–4.2
BV16	16	5	80	5	10	TSNFC	12	4
Broadband codecs								
G.722	48, 56, 64	0.0625	3–4	1.5	1.5625	SB-ADPCM	5	~4.1
G.722.1	24,32	20	480, 640	40	60	MLT	<15	~4
AMR-WB (G.722.2)	6.6-23.85	20	132-477	25	45	ACELP	≈ 38	Various
Speex (WB)	4-44.2	20	80-884	34	50	CELP	8-25	Various
iSAC	Variable 10–32	Adaptive 30–60 ms	Adaptive-variable	Frame + 3 ms	Adaptive 63–123	Transform coding	6–10	Various[e]
BV32	32	5	160	5	10	TSNFC	17.5	~4.1

[a] Every speech codec introduces a delay in the transmission. This delay amounts to the frame size, plus some amount of "look-ahead" required for examining future samples.
[b] The codec delay is the sum of the algorithmic delay and the time interval needed for processing purposes.
[c] MIPS stands for million instructions per second and represents a measure of a computer's processor speed.
[d] Theoretical maximum: 4.4.
[e] Better than G.722.2 at comparable bitrates.

`

RT_078, mes 2015
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba