RIELAC, Vol. XXXVI 3/2015 p. 56-68 Septiembre - Diciembre ISSN: 1815-5928

RIELAC
Revista de Ingeniería Electrónica, Automática y Comunicaciones

Aplicación de medidas de calidad en sistemas de reconocimiento de locutores

Claudia Bello Punto, Dayana Ribas González, Eniel Suárez Fernández, José R. Calvo de Lara

RESUMEN

En este trabajo se realiza un estudio acerca de la relación que existe entre medidas de calidad de la señal de voz y el comportamiento de un sistema de reconocimiento de locutores. Para ello se estudian las medidas de mayor utilidad en estos sistemas seleccionando cuatro de acuerdo con los parámetros que estas analizan en la señal y su importancia en el proceso de reconocimiento de locutores. Adicionalmente se analizan las diferentes variantes que existen para vincular la calidad a los sistemas de reconocimiento por lo que se llevan a cabo un conjunto de experimentos. Estos fueron realizados en una base masculina en varias condiciones de ruido aditivo para evaluar la relación entre el resultado del reconocimiento, la calidad de las muestras y el ruido presente en ellas a partir de la SNR. Fueron obtenidas conclusiones interesantes a partir de 1500 muestras y 20 escenarios de ruido diferentes.

Palabras claves: medidas de calidad de la voz, reconocimiento de locutores, ruido aditivo, kurtosis

Abstract

In this work, a study about the relationship between speech quality measuresand speaker recognition performance is presented. To accomplish this, the most significant quality measures for speaker recognition systems were analyzed and four of them (KLPC, KCEP, HD, P563) were selected according to the parameters considered to determine the quality and its relevance in the speaker recognition process. The alternatives for linking quality with speaker recognition performance are described and a set of experiments are conducted. Such experiments were performed in male database on several additive noise conditions to assess the relationship among the recognition result, the quality of the samples and noise present therein from the SNR. Interesting conclusions were obtained for 20 different noise scenarios.

Keywords: *Speech quality measures, speaker recognition, additive noise, kurtosis.*

Tittle: *Speech quality measures in speaker recognition systems.*

INTRODUCCIÓN

El estudio de la calidad en la voz se remonta a la década del 60 del siglo XX donde aparece la recomendación de la IEEE que define un primer método para determinar la calidad de una muestra [1]. Esta recomendación está referida amedir la calidad de manera subjetiva.Inicialmente el objetivo era medir el rendimiento del servicio de redes telefónicas. Los primeros métodos establecidos para realizar dicha medición se basaban en determinar la calidad de la voz a partir de la opinión de un conjunto de individuos.

Los resultados de estos métodos fueron eficaces por lo que se instauraron diversas recomendaciones [1 2 3] que definen la forma de aplicarlos de manera correcta.

El uso tan difundido de los métodos de procesamiento de la voz en aplicaciones de multimedia y telecomunicaciones eleva la necesidad de evaluar la calidad de las muestras de voz que se procesan. Por esta razón es necesario contar con una evaluación precisa y fiable de la calidad de la misma, que no solo satisfaga los requerimientos del usuario [4] sino que permita establecer un establecer un grado de confianza en los resultados obtenidos por el sistema.

El despliegue tecnológico alcanzado en las aplicaciones relacionadas con la voz ha sido muy amplio, tal es el caso de la telefonía celular, la transmisión de voz a través de redes IP yel reconocimiento del habla, del lenguaje y de locutores. En estos y en otros campos es preciso monitorear en tiempo real o determinar la calidad de la voz con una mayor exactitud, por lo que utilizar un grupo de expertos para determinar la calidad de la muestra no es factible [5].

Las condiciones inapropiadas en que se pueden adquirir las muestras de voz, traducidasen ruido que se mezcla con la señal original de diferentes maneras [6], pueden cambiar el comportamiento de estas en cualquier dominio de representación. La inteligibilidad es el factor principal si se trata de implantes cocleares o identificación de palabras claves donde se conoce que el ruido reduce la calidad de la muestra. En el caso específico de los sistemas de reconocimiento automático de locutores (SARL), la integridad de las características discriminativas del locutor es de mayor importancia, rasgos que también sufren modificaciones en presencia de ruido. Varios trabajos han resultado en métodos automáticos para medir la calidad, siendo totalmente independientes de la opinión de un individuo. A pesar de no haber sido concebidos para este tipo de sistemas se han utilizado, con resultados alentadores, en distintas etapas de un SARL para obtener resultados más exactos.

Este trabajo realiza la selección de cuatro métodos, a partir de los principales que se han encontrado en la literatura, con el fin de determinar la calidad de las muestras de voz y su relación con el comportamiento del sistema en varios escenarios ruidosos. Asimismo se describen las diferentes variantes que existen para vincular la calidad con un SARL y obtener un resultado más eficaz. Además de los conceptos generales sobre la calidad, se ofrece en el siguiente epígrafe una descripción de las principales medidas de calidad existentes, a partir de las cuales se realizó la selección. Luego se acotan los diferentes usos que puede tener la calidad en un SARL. Seguidamente se describen un conjunto de experimentos que permiten establecer la relación que existe entre la calidad de las muestras, el comportamiento del sistema y el ruido usando la Relación Señal Ruido*(Signal to Noise Ratio (SNR)).* Para finalizar se arriba a conclusiones a partir de los resultados obtenidos.

Medidas de calidad

Existen dos maneras de medir la calidad: de manera subjetiva donde un grupo de expertos escucha la muestra calificándola dentro de una escala predefinida, y de manera objetiva donde se obtiene un valor equivalente a la calidad de la muestra de manera automática. Las pruebas subjetivas son quizás los métodos más confiables para determinar la calidad, sin embargo requieren gran cantidad de tiempo y de recursos, por lo que no son apropiadas para aplicar en sistemas en los que se quiera, por ejemplo, determinar diariamente la calidad de servicio *(Quality of Service (QoS))* en una red IPa través de la cual se transmite voz (VoIP), o en un sistema de control de acceso a través de la voz donde es necesario evaluar la calidad de manera instantánea. Las medidas de calidad objetivas son capaces de realizar esta tarea de forma automatizada y a muy bajo costo. Por esta razón una gran parte de las investigaciones en este tema se han centrado en diseñar medidas objetivas para medir la calidad de las muestras de voz.

Evidentemente para que una medida de calidad objetiva sea válida es necesario que esté correlacionada de alguna forma con las medidas subjetivas, por esta razón varios métodos [7 8] se han encaminado a desarrollar medidas objetivas que modelen varios aspectos del sistema auditivo del hombre, que es el ejecutor de la medida subjetiva.

Las medidas objetivas se clasifican en intrusivas o no intrusivas en dependencia de requerir o no una muestra de voz original [9]. Las primeras proponen la aplicación de métodos de comparación entre la señal original y la degradada, y determinan la calidad cuantificando la diferencia entre ambas muestras. En este grupo se encuentran medidas basadas en la SNR [4 10], otras utilizan los coeficientes de predicción lineal *(Linear Prediction Coefficients (LPC))* [11 12] como base para medir la calidad, existen también un conjunto de medidas intrusivas que parten del modelo de percepción [7 8] y otras que combinan varias de las anteriores para evaluar una mayor cantidad de parámetros en la señal.

Sin embargo las medidas intrusivas tienen entre sus limitantes la necesidad de requerir la señal original.Esto es un gran inconveniente debido a que en algunas aplicaciones esta no está disponible. Si nos remitimos al ejemplo de VoIP donde se necesita monitorear continuamente el comportamiento de la red en un punto específico (en términos de la calidad de la voz) solamente se tiene acceso a la señal de salida. En este caso solamente una medida de calidad no intrusiva es adecuada para dicha tarea.

En SARL el comportamiento es bastante similar, sobre todo si se remite a aplicaciones forenses donde frecuentemente llegan grabaciones únicas de un individuo sin identificar.

Sucede también en otras aplicaciones como la autenticación del usuario o el manejo personalizado de datos a través de la voz, donde la señal que se tiene para realizar el reconocimiento es generalmente la que sale del procesamiento y por lo general no está limpia.
Por estas razones el enfoque para analizar una señal procesada sin que se cuente con su equivalente original es bien diferente por lo que se ha diseñado medidas objetivas no intrusivas. Algunas de ellas modelan el tracto vocal para calcular la distorsión, otras evalúan el comportamiento estadístico de la señal o de parámetros extraídos de ella para emitir un criterio de calidad. Un conjunto de medidas intenta estimar la calidad subjetiva de la muestra y otro grupo determina la SNR. Seguidamente se describen un conjunto de medidas objetivas no intrusivas.

Recomendación UIT-T P.563

En esta recomendación se describe un método objetivo no intrusivo para determinar la calidad subjetiva de la voz en aplicaciones de telefonía de 3.1kHz (banda estrecha). Se define como el criterio de calidad que brindaría un experto que está escuchando una llamada real con un teléfono convencional conectado en paralelo a la línea. El aporte que tiene esta recomendación es que es la primera que realiza mediciones no intrusivas que tiene en cuenta toda una gama de distorsiones que se producen en una red telefónica convencional, y permite predecir la calidad vocal sobre una escala de Puntuación Media de Opinión *(Mean Opinion Score (MOS))* de acuerdo con la Recomendación UIT-T P.800 [3].

Entre las condiciones para las que fue validada esta medida, devolviendo resultados aceptables, se encuentran el ruido ambiental en el lado de emisión, errores en el canal de transmisión, perdida de paquetes, transcodificaciones, deformaciones a corto y largo plazo de la señal,sistemas de transmisión con compensadores de eco y sistemas de reducción del ruido en condiciones de un solo hablante, entre otras. Sin embargo devuelve resultados inexactos cuando se trata del efecto del retardo en conversaciones, música o tonos de la red como señal de entrada.Este algoritmo no ha sido diseñado para aplicarlo al reconocimiento de locutores, sin embargo por los parámetros que mide en su evaluación de la calidad se ha demostrado su utilidad ya que abarca un gran número de condiciones que pueden estar presentes en este tipo de sistemas y que son propias de una señal procesada por él.

La señal se procesa de varias maneras, a modo de capas, que detectan un grupo de parámetros característicosde la señal. Sobre la base de un conjunto restringido de parámetros clave se asigna una clase de distorsión principal a la señal. Luego los parámetros clave y las clases se emplean para ajustar el modelo de calidad vocal que proporciona una ponderación perceptual, con la presencia de varias distorsiones sobre la señal y donde una clase predomina sobre el resto. La figura 1 resume el proceso por el que transita la señal hasta obtener un valor final de calidad en la escala MOS.

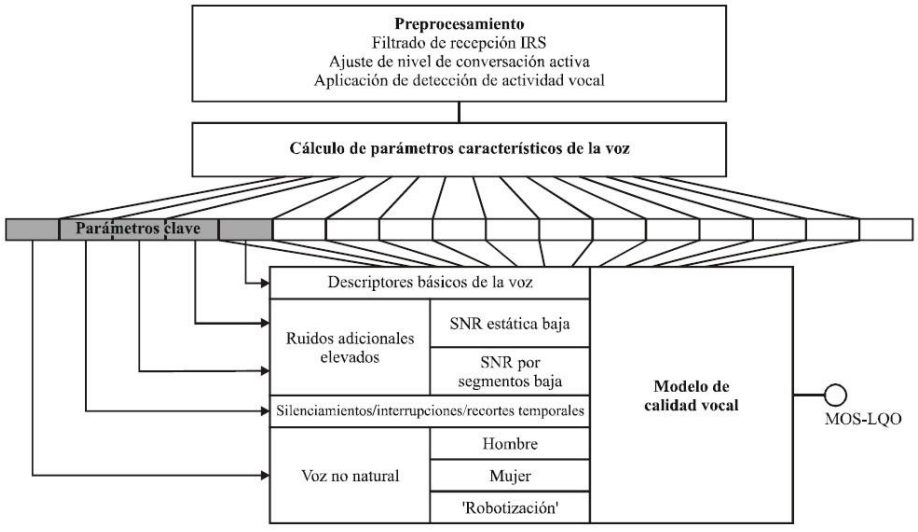

Fig. 1. Diagrama en bloques de UIT-T P.563 [13]

La parametrización de la señal se divide en tres bloques funcionales principales que se corresponden con las tres clases de distorsión: el primero incluye el análisis del tracto vocal y desnaturalización de la voz donde se analiza el género y robotización que pueda existir, el segundo profundiza en el análisis de ruido adicional intenso donde se determina la SNR estática reducida y la SNR por segmentos reducida, mientras que el tercero incluye las interrupciones, silenciamientos y el recorte temporal.

Análisis del tracto vocal y desnaturalización de la voz

Este bloque trata de detectar el carácter desnaturalizado de la voz, a partir de un modelo del tracto vocal, extrayendo partes de la señal que podrían interpretarse como voz y separarlas de las partes no vocales. Además ofrece información sobre la humanización de la voz a través de un análisis estadístico de manera diferenciada para voces masculinas y femeninas. Detecta además la presencia de tonos tales como los de marcación telefónica *(Dual-Tone Multi-Frequency signaling (DTMF))* o señales similares marcadamente periódicas no vocales. Se analizan las tramas de voz repetidas ocasionadas por la pérdida de paquetes en sistemas de transmisión en modo paquete.

Análisis del ruido adicional intenso

El análisis del ruido calcula distintas características del mismo. Este bloque se encarga de detectar si el ruido es la principal causa de degradación, si esto ocurre se analiza entonces el tipo de ruido que afecta la señal. Éste puede ser estático y estar presente en toda la señal (al menos durante la actividad vocal) de forma que la potencia de ruido no está correlacionada con la señal vocal, o bien, puede ocurrir que la potencia de ruido presente una cierta dependencia con respecto a la envolvente de la potencia de señal. Si se trata del primer caso entonces se realiza un análisis más específico sobre los fonemas y frases de la señal.

Interrupciones, silenciamientos y recorte temporal

Dichas distorsiones sólo pueden ser parcialmente descritas por el resultado del análisis del tracto vocal. Por tanto, se realiza nuevamente un análisis del tracto para detectar y valorar los recortes temporales y los silencios antinaturales. La interrupción de la señal puede ocurrir de dos formas diferentes, como un recorte temporal de la voz o como una interrupción de la misma. Ambos producen una pérdida de información de la señal. El recorte temporal puede ocurrir cuando se utiliza la detección de actividad vocal o se interrumpe la señal.

Distancia entre armónicos (HD)

HD calcula la razón de energía en las áreas más significativas para un SARL: armónicos de la frecuencia fundamental [(1)] y los valles entre ellos (Ec. 1). Este valor aumenta con la diferencia entre picos y valles, definiendo cuan limpia esta la señal. En las regiones espectrales afectadas por ruido, la energía aumenta rellenando los valles, lo que causa un decrecimiento de HD.
Para determinar la distorsión de la estructura armónica de la voz, esta medida define una función relacionando la potencia en los armónicos con la potencia en los valles.

$$HD = \frac{1}{NH * N_{FRAMES}} \sum_{N_{FRAMES}} \sum_{k=1}^{NH} 10 * \log\left(\frac{P_k + P_{k+1}}{2 * P_{k,k+1}}\right) (1)$$

Donde $P_k$es la potencia en el armónico k, $P_{k,k+1}$ es la potencia entre armónicos k y $k+1$. NH se refiere a la cantidad de armónicos y N_{frame} es el número de tramas de voz en la señal [14]. Es preciso destacar como esta medida de calidad solo se calcula en las tramas de voz y no en las de silencio pues son las que poseen la frecuencia fundamental y por tanto sus armónicos. Por esto es necesario determinar primero las zonas de voz en la señal. La razón de potencia definida en la Ec.1 aumenta con la SNR, por tal motivo se espera que al aumentar el ruido en la señal disminuya el valor de esta medida.

Kurtosis de Cepstral (KCEP)

La kurtosis o momento estadístico de cuarto orden es una medida de forma de la distribución de probabilidad de los valores reales de una variable aleatoria.

[(1)]Los armónicos de la voz son los que generan el timbre característico de la misma y a su vez permiten reconocer el timbre de la voz de una persona. Su frecuencia es un múltiplo de la frecuencia fundamental, que es la frecuencia a la que vibran las cuerdas vocales. Portan información discriminativa del locutor.

Con el incremento de la kurtosis el pico de la distribución crece, por tanto las muestras estarán más concentradas alrededor de la media. En este caso la kurtosis se aplica a la distribución de los Coeficientes Cepstrales en escala Mel *(Mel Frequency Cepstral Coefficients (MFCC))*. Esta medida de calidad utiliza la forma de la distribución de los MFCC como un indicador de degradación. En cada trama se obtienen P coeficientes, para luego determinar la kurtosis de la siguiente manera:

$$kcep = \frac{1}{P}\sum_{p=1}^{P}\left(\frac{c_p - \frac{1}{P}\sum_{p=1}^{P} C_p}{\sigma}\right)^4 \quad (2)$$

Donde P se refiere a la cantidad de coeficientes, C_pa los coeficiente MFCC y σ es la desviación estándar de la trama que se analiza. Para determinar $kcep$se eliminan todas las tramas de silencio usando un detector de actividad vocal. Posteriormente se promedian los valores de kurtosis de las tramas de voz para obtener el valor final de la medida de calidad.

El ruido aditivo causa un incremento de la energía en las componentes de la señal relativas al ruido, reflejándose en la densidad espectral de potencia *(Power Spectral Density (PSD))*y también en los MFCC. La figura 2 muestra la PSD de una trama de voz corrupta con ruido de exteriores para SNR=5 y 20 (este fenómeno se ilustra con una sola trama pero es similar en el resto de la señal). Para 20 dB, la PSD, tiene altos valores en las bajas frecuencias y decrece en las altas. Para 5 dB el incremento que se produce en la potencia de los segmentos de ruido implica mayores valores en PSD trayendo consigo cambios en la forma del espectro que se refleja en la distribución de los MFCC. Estos coeficientes se calculan utilizando la Inversa de la Transformada de Coseno Discreta*(Inverse Discrete Cosine Transform (IDCT))*, los cuales expresan la forma de la señal en funciones coseno [15]. Los coeficientes reflejan la similitud de la señal con las funciones base. De esa manera, C_1 representa la disminución de la pendiente de la señal, C_2 será más negativo indicando la tendencia a un ciclo de coseno y C_3 aumentará representando el comportamiento de un ciclo de coseno desplazado 90 grados. En consecuencia la distribución de los MFCC esmás dispersa, por lo que disminuirá $kcep$ en señales ruidosas (como indican los valores en la figura 2).

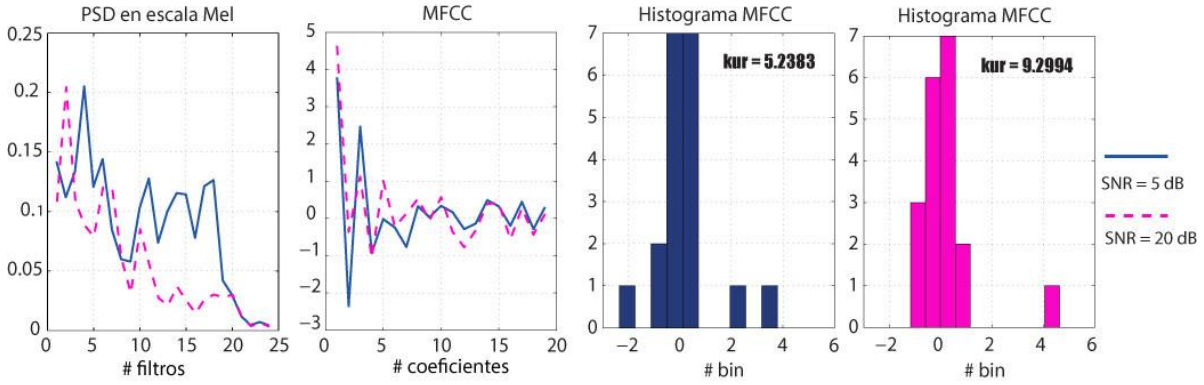

Fig. 2 Segmento sonoro mezclado con ruido de exteriores a SNR=5 y 20 dB.

Kurtosis sobre los LPC

La kurtosis sobre los LPC es similar a la medida descrita en el apartado anterior, solo que esta utiliza la forma de la distribución de los rasgos LPC como indicador de degradación de la muestra de voz. Se mantiene por tanto la misma ecuación, solo se sustituye C_p por a_p, siendo este término el coeficiente LP que se analiza. El análisis LP consiste en estimar el modelo del tracto vocal partiendo de muestras previas. Cuando la muestra es parte de un segmento que sigue determinado patrón, tal es el caso de un segmento periódico, solo se necesita un pequeña referencia para predecir el comportamiento de la muestra. Este es el caso de una trama de voz, cuando el predictor está situado en zonas sonoras que son cuasi periódicas, la predicción se realizará prácticamente en las zonas cercanas, especialmente en las muestras previas. Por tanto, el primer coeficiente (a_2) tendrá valores cercanos a -1 mientras que el resto tendrán poca influencia en la predicción, con valores muy cercanos a 0. En consecuencia se obtendrá una distribución con valores altamente concentrados alrededor de la media y por tanto altos valores de kurtosis. Si la señal es ruidosa, perderá periodicidad y las muestras por tanto no serán tan predecibles.

Por este motivo todos los a_p participan en la predicción, tomarán valores diferentes y alejados de cero lo que produce un esparcimiento en la distribución de los LPC y menores valores de kurtosis.

Criterio UBML

Es una medida basada en modelos estadísticos que aprovecha los modelos de habla poco degradada para determinar la calidad. Esta aproxima la similitud entre una locución y el modelo universal utilizado para generar el modelo estadístico de un locutor. Se obtendrá de manera inmediata si se utiliza un sistema basado en Modelos de Mezclas de Gaussianas *(Gaussian Mixture Model (GMM)*, ya que para determinar la puntuación de la similitud es necesario calcular la verosimilitud entre el Modelo Universal de *background (Universal background Model (UBM))* y la locución de prueba (3).

$$S(O,\lambda_t) = \log p(O,\lambda_t) - \log p(O,\lambda_{UBM}) \tag{3}$$

Donde λ_t representa el modelo GMM del locutor y O los rasgos extraídos de la locución.Esta medida de calidad se basa en la idea de que si un UBM está entrenado bajo determinadas condiciones, una locución con características diferentes tendrá un peor comportamiento porque el UBM no le es representativo y por tanto se le debe asociar una calidad baja, así esta medida es una idea de lo diferentes que son las muestras que se utilizan en un sistema de reconocimiento con respecto a las utilizadas para entrenar el mismo [5]. La medida se determina de la siguiente manera:

$$UBML = \log p(O,\lambda_{UBM}) \tag{4}$$

Donde $p(.,\lambda_{UBM})$ es la función densidad de probabilidad para cualquier modelo λ.

SNR

La SNR cuantifica en qué medida una señal $x(t)$ ha sido afectada por un ruido $n(t)$. En este caso la señal es la voz y el ruido corresponde a una perturbación acústica-aditiva, según el siguiente modelo:

$$y(t) = x(t) + n(t) \tag{5}$$

Luego la SNR es la razón de potencias entre la voz y el ruido de fondo, definida como sigue:

$$SNR(db) = 20\log_{10}\frac{Pvoz}{Pruido} \tag{6}$$

$$Px = \sqrt{\frac{1}{T}\sum x(t)^2} \tag{7}$$

donde Px es la potencia de voz o ruido, $x(t)$son las muestras de voz en el tiempo y T se refiere a la cantidad de muestras de voz. En la ecuación (6), se observa que la SNR es inversamente proporcional a la potencia relativa al ruido $Pruido$, por lo tanto a mayor de SNR menor será la variabilidad de la señal y mayor la calidad de la misma. Al ser esta variante de SNR no intrusiva solo se cuenta con $y(t)$ por lo que es preciso estimar $n(t)$ para luego determinar la SNR. Por este motivo la eficacia del cálculo estará en dependencia de cuan preciso sea el método de estimación utilizado. Su selección debe estar en correspondencia con el tipo de ruido que afecte la señal.

Uso de medidas de calidad en SARL

Ninguna de las medidas descritas anteriormente, a excepción de UBML, fueron diseñadas para usarse en SARL. Sin embargo, varios autores han vinculado la calidad al resultado de estos sistemas desde diferentes enfoques. En Kelly *et all* [16] se relacionan el envejecimiento, la calidad y el resultado de la verificación con el objetivo de observar la influencia que tienen ambos factores en la tarea de reconocimiento de locutores. Es posible utilizar también la calidad de la señal en varias etapas del sistema, por ejemplo, en las etapas de extracción de rasgos, entrenamiento de los modelos, determinación de la puntuación y fusión de estas. En García-Romero [17] han mostrado resultados alentadores cuando se incorpora la calidad en el proceso de reconocimiento sobre todo en las dos últimas etapas. A continuación se muestra las aplicaciones más relevantes que ha tenido la calidad en SARL.

Medidas de calidad aplicadas durante el cálculo, la fusión y la calibración de *scores*

En este epígrafe se presentan varios métodos para vincular la calidad de una muestra al cálculo y la fusión de scores utilizando diferentes niveles de información en la señal de voz debido a la clara relación que existen entre la puntuación resultante y dicha información dado que porta información discriminativa del locutor (IDL). La idea parte de que los seres humanos mezclan varios niveles de información para reconocer la identidad de un locutor.

Usualmente para el cálculo de la puntuación se utiliza un sistema GMM-UBM, utilizando una etapa de pre-procesamiento en las que se eliminan los efectos del canal y se reduce el ruido presente en la muestra. Además se eliminan los silencios y los sonidos que no se consideran voz, preservando solo la información que satisfaga determinado criterio, eliminando el resto.

Si se combina esta etapa con un mecanismo clásico para determinar la puntuación, se le confiere a toda la información que sepreserva, luego de la etapa de pre procesamiento, la misma importancia. Sin embargo se omite que al utilizarla para determinar el por ciento de verificación no se tiene en cuenta que la información referente al locutor y la que puede degradar la muestra no están distribuidas de manera uniforme en la señal. Si la puntuación se calcula incluyendo la calidad esta actúa como un factor de peso en dicha etapa.

Este concepto se puede aplicar a cualquier técnica usada en sistemas de reconocimiento, pero en este caso se particulariza en GMM para nivel espectral ya que es el más utilizado en la literatura. El procedimiento se modifica quedando de la siguiente manera:

Dado una secuencia de vectores de rasgos $O = \{o_1, o_2, o_3, \dots o_T\}$, donde T corresponde a la cantidad de trama de la señal a la que se aplica la medida de calidad ξ. La probabilidad del modelo λ incorporando el valor de calidad como un factor de peso se determina de la siguiente manera:

$$p(O|Q,\lambda) = \prod_{t=1}^{T} p(o_t|\lambda)^{q^{\xi}} \tag{8}$$

Luego el logaritmo de la probabilidad se determina como:

$$\log p(O|Q,\lambda) = \sum_{t=1}^{T} q_t^{\xi} \log p(o_t|\lambda) \tag{9}$$

Para incluir varios niveles de información, es necesario acudir a la fusión y se utiliza una Máquina de Soporte Vectorialadaptada para poder incluir la información relativa a la calidad en el proceso de verificación. El método se basa en la combinación de información de bajo nivel (por ejemplo, información espectral) con otros tipos de información de alto nivel (información fonética y lexical). Esta idea parte de que los sistemas de verificación que utilizan información de bajo nivel tienen mejores resultados que los que utilizan información de alto nivel. Además se basa en que las afectaciones que se producen en el primer caso son más fáciles de detectar que en el segundo, por lo que el diseño de las medidas de calidad será más sencillo para los sistemas que utilizan bajos niveles de información. A partir de ello se propone un sistema que utiliza la calidad como el factor de decisión para definir si usar un SARL solamente basado en información espectral o si combina con un sistema que utiliza información de alto nivel para determinar la puntuación final

A partir de esta idea se propone un modelo llamado Fusión de *Scores* Basado en la Calidad donde la información de calidad se incorpora como un factor de decisión para utilizar el sistema solo con el mejor comportamiento, es decir basado en información de bajo nivel, o combinando ambos sistemas. Esta modificación implica que la puntuación resultante será al menos tan exacta como la del sistema de mejor desempeño, o mejor.

La calidad también puede ser útil para calibrar el SARL. En [18] se utiliza la duración de las señales como una medida de calidad para calibrar el sistema debido a que la variabilidad de esta conduce a la disminución de su rendimiento. El sistema de verificación utilizado se basa en *i-vector*, pues se ha demostrado que este es menos sensible a las muestras de poca duración comparada con sistemas basados en SVM y Análisis de Factores. En este caso se utiliza la duración de los segmentos de entrenamiento y prueba como medida de calidad para calibrar el sistema (11). Esta técnica de calibración se conoce con el nombre de (*Quality Measure Function (QMF)*).

La duración es solo un ejemplo de medida de calidad, esta manera de realizar la calibración de un sistema puede basarse en cualquier otra medida.

La calidad de la voz permite predecir tanto el rendimiento de un sistema de verificación de locutor como una posible desalineación de los *scores* del mismo debido a cambios en dicha calidad.

Los *scores* de los clientes e impostores pueden desajustarse y la calidad es una variante para realizar un ajuste de esta diferencia.Puede recurrirse a graficas de dispersión, curvas DET para determinar estas variaciones. Otra manera de hacer este análisis es evaluar el aporte que tiene la calidad en el resultado, para ello se define en Castro [5] una medida del impacto que tiene la calidad en el score (10).

$$Impacto = \frac{EER_{max} - EER_{min}}{EER_{max}} \tag{10}$$

donde $\mathrm{EER_{max}}$ se refiere al mayor valor de *score* obtenido para una medida de calidad determinada mientras que $\mathrm{EER_{min}}$ corresponde con el mínimo. Es necesario destacar que este valor solo da una idea de utilidad de manera parcial por lo que se sugiere analizar además las curvas de EER vs valor de la medida de calidad. Aquí además se incluye en el análisis la calidad de la muestra utilizada para el entrenamiento del sistema definiendo la calidad de la comparación como:

$$Q = \sqrt{Q_{train} Q_{test}} \tag{11}$$

donde Q_{train} y Q_{test} corresponden a la calidad de las muestras para entrenamiento y prueba.

Medidas de calidad para verificar la confiabilidad en la decisión de un sistema de reconocimiento de locutor.

Esta variante de aplicación tiene como objetivo determinar cuan confiable es la decisión tomada por un SARL una vez ejecutada la tarea. Las herramientas utilizadas con mayor frecuencia son las redes bayesianas y las redes neuronales.
El uso de las redes bayesianas esta aplicación fue propuesto por primera vez en [19] [20] [21] debido a que estos modelos gráficos permiten determinar la probabilidad de la confiabilidad de la decisión.
La idea inicial en estos trabajos es elevar el rendimiento del sistema solicitando al usuario una nueva muestra si la que se analiza no tiene suficiente calidad. No se desecha ninguna de las muestras tomadas, cada score obtenido con ellas se pesa con el valor de la confiabilidad que devuelve la red y así obtener el resultado final S_c.

$$S_c = \sum_n Conf(Sc_n)\, Sc_n \tag{12}$$

Donde Sc_n es el score obtenido para la muestra ny$Conf(Sc_n)$es la confianza devuelta por la red bayesiana.
Originalmente no se contemplaba en la arquitectura de la red una relación entre la calidad y la clase a la que pertenecía la comparación (cliente o impostor), cuando ya en [6] se había probado que la calidad afectaba de manera diferenciada a ambas clases. Es por esto que en [22] se incluye esta relación y a pesar de que la confiabilidad se utiliza para eliminar comparaciones no confiables los resultados superan la primera propuesta comprobando que la distribución de los *scores*se afecta de manera diferenciada para clientes e impostores. Luego en [23] se propone una modificación a la arquitectura de [22] esta vez eliminado la confiabilidad de los nodos de la red y obteniendo una puntuación limpia que corresponde con el valor que se hubiera obtenido si la muestra tuviera buena calidad. Con este score limpio se define luego una regla para determinar la confiabilidad de la decisión y descartar las comparaciones no confiables. Los resultados obtenidos superan las propuestas antes descritas. En [22] [23] se incluye en el análisis la calidad referente a las muestras usadas en la fase de entrenamiento y de prueba.
Las redes neuronales han sido mucho menos usadas con este fin, solamente en [24] se trabaja en este tema, sin embargo estos métodos tienen como gran desventaja la gran cantidad de comparaciones que son necesarias para poder realizar el entrenamiento de la red y por tanto el alto costo computacional que esto conlleva.

EXPERIMENTOS

Dado que el objetivo de ese trabajo es evaluar la relación que existe entre la calidad, el resultado del sistema y la cantidad de ruido presente en las muestras se elige la primera variante descrita en el apartado anterior haciendo un análisis de correlación como se describe más adelante.

Selección de las medidas de calidad

De los grupos que forman el conjunto de medidas de calidad objetivas no intrusivas se seleccionaron cuatro medidas con el fin de evaluar su relación con el ruido y con la puntuación de un SARL.La P.563 se selecciona debido a que evalúa una gran cantidad de propiedades de la señal para arribar a un resultado final (MOS). HD fue seleccionada del grupo de métodos que estiman el ruido presente en la señal. Esta medida determina la SNR a partir de la frecuencia fundamental, que porta información discriminativa del locutor, de ahí la importancia de su selección. Finalmente, del grupo de medidas estadísticas la KCEP y la KLPC fueron elegidas pues se basan en parámetros que usualmente se utilizan para realizar el reconocimiento.

Corpus

Para llevar a cabo los experimentos se utilizaron 50 locutores de la base NIST 2001 en idioma español. Se tomaron 50 muestras de una sesión microfónica para el entrenamiento y 50 de otra sesión microfónica para la prueba. Para crear las muestras ruidosas, se mezclaron de manera electrónica las muestras correspondientes a esta fase con: ruido blanco estacionario, ruido de exteriores pseudo-estacionario, ruido de voces no estacionario, el cual está altamente correlacionado con la voz debido a que se compone de voces de diferentes locutores y ruido de música no estacionario y altamente armónico, en 5 niveles de SNR [25]. Los detalles del SARL utilizado se pueden localizar en [26].

RESULTADOS Y DISCUSIÓN

Calidad vs. SNR

Esta relaciónse evalúa con el objetivo de analizar cuan exactas son las medidas de calidad seleccionadas reflejando la cantidad de ruido presente en la señal. Mayores valores de SNR implican menos ruido en las muestras y por lo tanto mayor calidad. La figura 3 muestra una media de los valores de calidad en los diferentes entornos ruidosos.

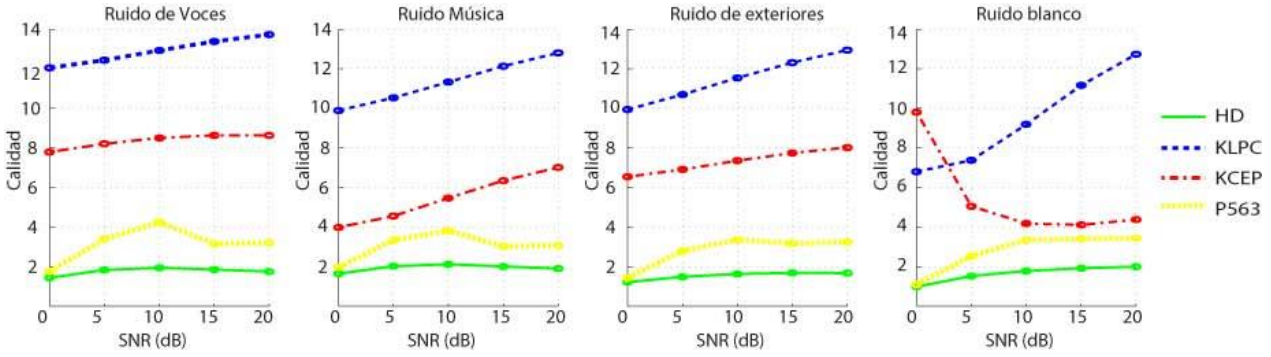

Fig. 3 Calidad vs. SNR

HD se encuentra representada por la curvas de color verde. Como es posible observar estas crecen cuando se trata de ruido blanco y de exteriores, mostrando como la medida refleja la cantidad de ruido en la señal. Este comportamiento se debe a que son ruidos con características muy estacionarias por lo que afectan la señal de manera uniforme y la medida los detecta correctamente. Principalmente en presencia de ruido blanco, totalmente estacionario, la energía se modifica de manera constante en todas las regiones, por lo que la razón entre armónicos y valles va a disminuir mostrando así como disminuye la calidad con la presencia de ruido.

Sin embargo aparece un cambio de pendiente alrededor de los 15 dB en los casos de ruido de voces y música, debido a que las componentes espectrales de estos tipos de ruidos no son uniformes, es decir, no se presentan durante toda la señal de manera constante.

En estos escenarios, algunos armónicos y valles se encuentran más corruptos por ruidos que otros por lo que se produce distorsión en el resultado de la medida.

La P.563, descrita por las curvas amarillas, aumenta con la SNR si se trata de entornos en los que aparece el ruido blanco, en caso contrario las curvas cambian bruscamente en los 15 y los 20 dB. Este comportamiento se debe a que esta medida cuenta con un bloque de detección de ruido, que devuelve valores incorrectos en muestras con altos valores de SNR, como usualmente ocurre con los métodos de detección de ruido ante señales limpias, lo que trae consigo una estimación incorrecta de la calidad en estos casos.

Las líneas azules muestran el comportamiento de la KLPC y se observa que estas aumentan con la SNR en todos los escenarios, demostrando que es una medida muy consecuente con los niveles de ruido que se encuentran en las señales, mostrando el mejor resultado en relación con el resto de las medidas seleccionadas. Es necesario destacar que para todos los tipos de ruido en SNR=20 dB las curvas alcanzan diferentes valores de kurtosis, lo cual es relativo a las características propias del ruido y a la manera en que este afecta la señal.

La KCEP, descrita por las líneas rojas, refleja correctamente el ruido en todos los escenarios excepto para ruido blanco, en el que tiene un comportamiento totalmente inverso. Esto se debe a que este tipo de ruido presenta energía en todas las componentes de frecuencia y al mezclarse con la señal produce un suavizado del espectro, es decir los valles entre los armónicos se rellenan. A consecuencia de esto se produce un incremento considerable en los filtros de alta frecuencia que se refleja en la PSD debido al ancho de banda logarítmico de este banco. Por tal motivo se reduce el coeficiente C_1, a causa de la inversión que sufre la pendiente de la PSD en escala Mel. En el histograma se distingue una concentración alta alrededor de 0, enfatizada por la gran diferencia que impone C_1 en relación con el resto de los coeficientes, por lo que la KCEP aumentara en estas circunstancias.

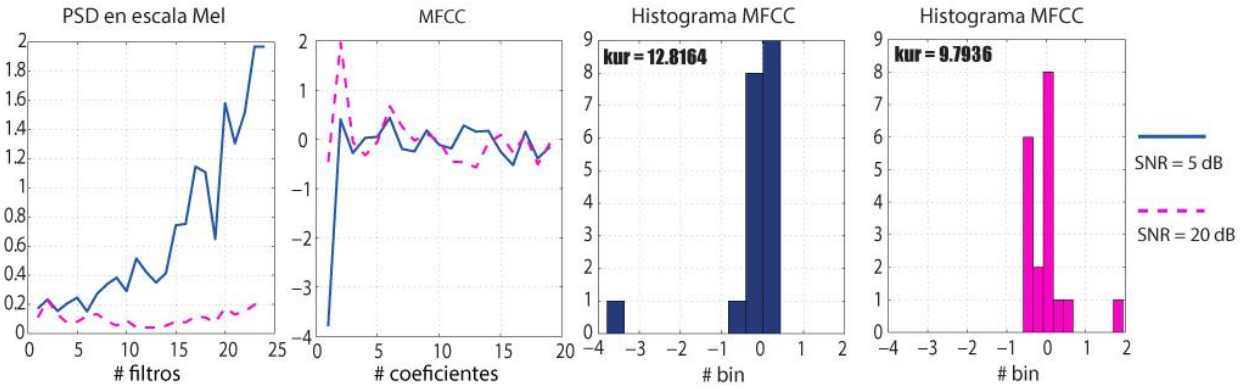

Fig. 4 Segmento sonoro mezclado con ruido blanco a SNR = 5y 20 dB.

Calidad vs. Puntuación del sistema (*Score)*

Este experimento se realizó con el objetivo de evaluar qué relación existe entre la calidad y el resultado de un SARL, utilizando el *score.* Para cada ruido y nivel de SNR se realizó un análisis de correlación entre ambos parámetros obteniendo el coeficiente de correlación lineal *R* entre ellos. Las columnas 3 y 6 de la tabla 1 muestran un promedio sobre los resultados para cada nivel de SNR.

Se supone que al mejorar la calidad de las muestras la puntuación que devuelva el sistema al realizar la verificación deba incrementarse y por tanto se obtengan valores positivos de R. Sin embargo, solo se obtienen valores muy cercanos a 0 e incluso negativos en la mayoría de las condiciones, lo que indica baja correlación entre la calidad y el *score.* De acá se puede inferir que el ruido no es el único factor que afecta la eficacia de los SARL. La diferencia de sesión es también un factor de peso en este resultado [27], incluyendo además la diferencia entre los dispositivos de adquisición de las muestras, los efectos causados por el canal, el estado de ánimo del locutor, la fonética, etc. [28]. Además la información discriminativa del locutor (IDL) también tiene un papel fundamental en la decisión.

A partir de estas observaciones, los valores de R se calcularon nuevamente, esta vez eliminando la diferencia de sesión.

Para ello se aisló el ruido aditivo utilizando las mismas muestras para entrenar y luego realizar la prueba, además, los scores fueron normalizados con respecto a las muestras limpias. Las columnas 4 y 7 de la tabla 1 muestran un promedio sobre la SNR para estos nuevos resultados.

Ahora *R* es mayor, indicando un incremento en la relación lineal que existe entre la calidad y el *score* en la mayoría de los casos analizados. Es de notar también como se incrementa el número de valores positivos, indicando que los valores tan bajos obtenidos en el primer experimento son relativos al *score* como medida del SARL pues además del ruido aditivo, este se afecta por otros elementos relativos a las características de la comparación y de las muestras. Por lo tanto estas dos medidas no pueden relacionarse de manera directa. Sin embargo, los resultados muestran que la KCEP y la P.563 son las medidas que guardan mayor relación con el score dado que tienen los valores más elevados de *R*. Las nubes que se observan en la figura 5 corresponden a KCEP en caso de ruido de exteriores y en ella es evidente que cuando no existe diferencia de sesión existe menos dispersión y se aprecia una tendencia a crecer con la SNR. Mientras que sucede lo contrario cuando no coinciden las sesiones.

Ruido	Medidas de calidad	Diferencia en las sesiones	Coincidencia en las sesiones	Medidas de calidad	Diferencia en las sesiones	Coincidencia en las sesiones
Voces	P.563	0.008	0,255	KLPC	-0,148	0,071
Música		-0.016	0,300		0,058	-0,129
Exteriores		0.141	0,316		0,009	-0,024
Blanco		0.108	0,341		0,057	-0,182
Voces	KCEP	-0,155	-0,01	HD	0,014	-0,022
Música		-0,053	0,234		-0,197	0,207
Exteriores		-0,645	0,309		-0,042	0,087
Blanco		0,067	0,295		-0,155	0,066

Tabla 1: Promedio del coeficiente de correlación entre el resultado del SARL y los valores de calidad

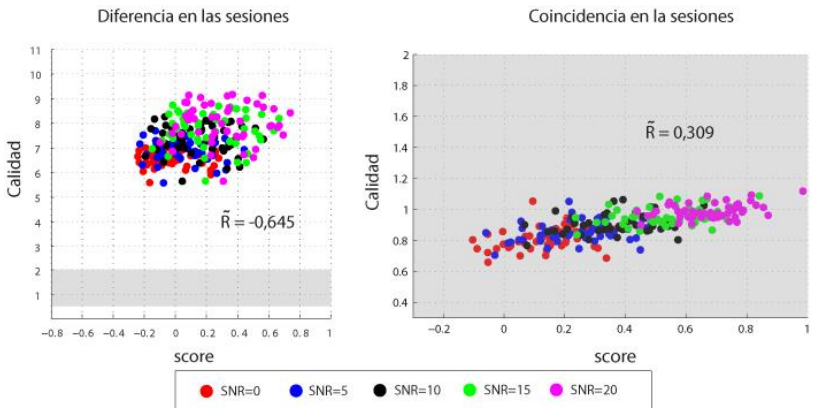

Fig. 5. Gráfico de dispersión de Calidad vs. Score para KCEP mezclada con ruido de exteriores para diferentes e iguales sesiones.

CONCLUSIONES

Este trabajo presenta un estudio sobre la relación entre la calidad de las muestras de voz y los resultados obtenidos de un SARL en ambientes ruidosos y para ello se utilizaron los valores de calidad y *score* de las muestras en diversas condiciones de ruido. Partiendo del comportamiento de todas las medidas analizadas los resultados muestran que: HD y P.563 reflejan el ruido en condiciones altamente difíciles (SNR<= 10 dB). No sucederá así para SNR>=15 dB principalmente para ruido de voces y música, mientras que las medidas estadísticas tienen buen comportamiento en todos los escenarios, excepto KCEP en presencia de ruido blanco, en cuyo caso el comportamiento es inverso.

Sin embargo este tipo de ruido no se encuentra comúnmente en aplicaciones reales, por lo que se recomienda elegir las medidas basadas en la kurtosis para detectar los niveles de ruido, especialmente KLPC, mientras que HD y P.563 para aplicaciones situadas en ambientes altamente ruidosos.

Por otro lado los resultados de correlación obtenidos se consideran moderados o bajos, mostrando que la relación entre la calidad y el score no es exactamente lineal. Estos experimentos permiten concluir que el resultado de los SARL está muy relacionado con la sesión y con la IDL además del ruido.

En esta evaluación se eliminó la diferencia de sesión, por lo que el siguiente paso estará encaminado a aislar los efectos que tiene la IDL en el resultado del SARL. Sin embargo en este momento pueden usarse estas medidas para tener una noción de los resultados del sistema. En el futuro se pretende llevar a cabo una extensión de este análisis a una base de datos mayor así como a SARL más actuales.

REFERENCIAS

1. IEEE Recommended Pratice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, **17**(3): p. 225-246, 1969.
2. **ITU-T Rec. P.830**. Calidad de la Transmision telefonica. Prueba subjetiva de opinion, en Sector de normalizacion de las telecomunicaciones,1998.
3. **ITU-T Rec. P.800**, Methods for subjectivs determination of transmission quality, en Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales,1996.
4. **Benesty, J., M.M. Sondhi, and Y. Huang**, *Springer Handbook of Speech Processing*, Springer-Verlag New York, Inc, 2007.
5. **Castro, A.H**., Fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz, en Dpto. de Ingeniería Informática, Universidad Autónoma de Madrid, 2010.
6. **Richiardi, J. and A. Drygajlo**. Evaluation of speech quality measures for the purpose of speaker verification. en *Proc. Odyssey: The Speaker and Language Recognition Worksho*p, 2008.
7. **Wang, S., A. Sekey, and A. Gersho**, An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, **10**(5): p. 819-829, 1992.
8. **ITU-T Rec. P.862**, Evaluación de la calidad vocal por percepción: Un método objetivo para la evaluación de la calidad vocal de extremo a extremo de redes telefónicas de banda estrecha y códecs vocales, en Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales, 2001.
9. **Kondo, K**., Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications. Springer, 2012.
10. **Loizou, P.C.**, Speech Quality Asssessment, en Multimedia Analysis, Processing and Communications, Springer, 2011.
11. **Kitawaki, N., H. Nagabuchi, and K. Itoh**, Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications*, **6**(2): p. 242-248, 1988.
12. **Itakura, F. and T. Umezaki**. Distance measure for speech recognition based on the smoothed group delay spectrum, en *Acoustics, Speech, and Signal Processing, IEEE International Conference* on ICASSP1987.
13. **ITU-T Rec. P.563**, Método basado en un solo extremo para la evaluación objetiva de la calidad vocal en aplicaciones de telefonía de banda estrecha, en Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales. Aparatos para mediciones objetivas. 2004.
14. **Vaseghi, S.V**., *Advanced Digital Signal Processing and Noise Reduction*. 4th ed.: John Wiley&Sons,2008.
15. Britanak, V., P.C. Yip, and K.R. Rao, *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations*. Academic Press, 2010.
16. **Kelly, F., A. Drygajlo**, **and N. Harte**, Compensating for Ageing and Quality variation in Speaker Verification, en *Interspeech* 2012.
17. **Garcia-Romero, D**., **et al**., Using quality measures for multilevel speaker recognition. *Computer Speech & Language*, **20**(2–3): p. 192-209, 2006.
18. **Mandasari, M**., **et al**., Quality Measure Functions for Calibration of Speaker Recognition System in Various Duration Conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, PP(99): p. 1-1, 2013.
19. **Richiardi, J., P. Prodanov**, **and A. Drygajlo**. Speaker Verification with Confidence and Reliability Measures. en *Acoustics, Speech and Signal Processing*, ICASSP 2006.

20. **Richiardi, J., P. Prodanov, and A. Drygajlo**, A probabilistic measure of modality reliability in speaker verification. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Pa, USA. **1**: p. 709 - 712, March 2005.
21. **Richiardi, J., A. Drygajlo, and P. Prodanov**, Confidence and reliability measures in speaker verification. *Journal of the Franklin Institute*, 343(6): p. 574-595, 2006.
22. **Villalba, J., et al**., *Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures, en Advances in Speech and Language Technologies for Iberian Languages*, Springer Berlin Heidelberg. p. 1-10, 2012.
23. **López, J.V., et al.**, A new Bayesian network to assess the reliability of speaker verification decisions, p. 3132-3136, en *Interspeech*. 2013.
24. **Campbell, W.M., et al.** Estimating and Evaluating Confidence for Forensic Speaker Recognition, en Acoustics, Speech, and Signal Processing, ICASSP 2005.
25. **Varga, A. and H.J.M. Steeneken,** Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,**12**(3): p. 247-251, *Speech Communication*, 1993.
26. **Ribas González, D. and J.R. Calvo de Lara,** Feature classification criterion for missing features mask estimation in robust speaker recognition,8(2): p. 365-375, *Signal, Image and Video Processing*, 2014.
27. **Ming, J., et al.**, Robust Speaker Recognition in Noisy Conditions. *IEEE Transactions on Audio, Speech & Language Processing*, 15(5): p. 1711-1723, 2007.
28. **Bimbot, F., et al.**, A Tutorial on Text-Independent Speaker Verification. EURASIP *Journal on Advances in Signa*l *Processing*, 2004(4): p. 101962, 2004.

AUTORES

Claudia Bello Punto, *Ingeniera en Telecomunicaciones y Electrónica, Reserva Científica, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Habana, Cuba, cbello@cenatav.co.cu.*

Dayana Ribas González, *Ingeniero en Telecomunicaciones y Electrónica, Master en Señales y Sistemas, Investigador Agregado, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Habana, Cuba, dribas@cenatav.co.cu.*

Eniel Suárez Fernández, *Estudiante de Ingeniería Biomédica, Universidad Central Martha Abreu de las Villas (UCLV), Santa Clara, Cuba***.**

José Ramón Calvo de Lara, *Ingeniero en Telecomunicaciones y Electrónica, Doctor en Ciencias Técnicas, Investigador Titular, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Habana, Cuba, jcalvo@cenatav.co.cu.*