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

# Medidas de calidad de la señal de voz aplicadas al reconocimiento de locutores

**Claudia Bello Punto, Dayana Ribas González y José R. Calvo de Lara**

**RT_058** **enero 2014**

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

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

REPORTE TÉCNICO
**Reconocimiento de Patrones**

# Medidas de calidad de la señal de voz aplicadas al reconocimiento de locutores

**Claudia Bello Punto, Dayana Ribas González y José R. Calvo de Lara**

**RT_058** **enero 2014**

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

# Medidas de calidad de la señal de voz aplicadas al reconocimiento de locutores

Claudia Bello Punto, Dayana Ribas González y José R. Calvo de Lara

Departamento de Biometría, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
{cbello, dribas, jcalvo}@cenatav.co.cu

RT_058, Serie Azul, CENATAV
Aceptado: 12 de diciembre de 2013

**Resumen.** La calidad de la muestra de señal de voz tiene gran influencia en un sistema de reconocimiento de locutores, pues condiciona el resultado obtenido en dependencia de la degradación presente en la señal. En este trabajo se abordan las principales medidas de calidad encontradas en el estado del arte. Además se hace referencia a la aplicabilidad de dichas medidas para determinar la veracidad de los resultados obtenidos en una comparación, así como para compensar la disminución del rendimiento de un sistema de reconocimiento de locutores. Para corroborar la relación entre la calidad de la señal corruptas por diferentes tipos de ruido, la relación señal ruido y el resultado del reconocimiento se evalúan cuatro medidas de calidad, KLPC, KCEP, HD y P.563. Resultados preliminares arrojan un comportamiento inverso de las dos primeras medidas, mientras que las últimas varían su desempeño en dependencia del tipo de ruido que se analiza.

**Palabras clave:** medidas de calidad de la voz, reconocimiento de locutores.

**Abstract.** The quality of a speech sample have great influence on speaker recognition system, it conditions the result obtained in dependence of the degradation in the signal. This paper provides an overview of the various methods used for assessment of speech quality. Also refers to the applicability of these measures, to determine the accuracy of the scores obtained in a comparison, as well as to balance the decline in performance of a speaker recognition system. In this work, we evaluate the performance of four quality measures, KLPC, KCEP, HD and ITU-T P.563 to study the close relationship between the quality of the speech signal corrupted by various types of noise, the signal to noise ratio and the speaker's recognition result. Preliminary experiments dump an inverse behavior for two formers quality measures while the performance of the lasts quality measures depend of the type of noise that degrade the signal

**Keywords:** speech quality measures, speaker recognition.

## 1 Introducción

El estudio de la calidad se remonta a la década del 60 del siglo XX donde parece la primera recomendación del Sector de Normalización de las Telecomunicaciones de la Unión Internacional de Telecomunicaciones, UIT-T [1]. Esta recomendación está referida de medir la calidad de una muestra

de manera subjetiva. Inicialmente el objetivo era medir el rendimiento del servicio de redes telefónicas. Los primeros métodos establecidos para realizar dicha medición se basan 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 [1-3] 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 del sistema [4] sino que permita establecer un grado de confianza en los resultados obtenidos por el método.

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 y el reconocimiento del habla, del lenguaje y de locutores. En estas y en otras aplicaciones 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]. En el caso específico del reconocimiento automático de locutores, una de las causas que degradan la calidad de la señal está relacionada con que las condiciones de adquisición de las muestras no sean las más adecuadas, lo que puede traducirse en ruido que se mezcla con la señal original de diferentes maneras [6]. Trabajos previos han resultado en diversos métodos que son totalmente independientes de la opinión de un individuo automatizando la medición de la calidad. En este trabajo se presenta un estudio sobre los métodos existentes, sus características, ventajas y limitantes, y se proponen probables líneas de investigación con el objetivo de obtener nuevos métodos para la medición de la calidad de la voz en el reconocimiento de locutores. Este trabajo está estructurado de la siguiente manera: secciones 2 y 3 se tratan los conceptos básicos relacionados al tema, los diferentes tipos de medidas de calidad que aparecen en la literatura. En la sección 4 se describen cuatro de las principales medidas vinculadas al reconocimiento de locutores mientras que en la sección 5 describen los posibles usos de las medidas de calidad en un sistema de reconocimiento de locutores. La sección 6 describe un conjunto de experimentos mientras que en la sección 7 se arriba a conclusiones y en la 8 se listan una serie de recomendación y una posible línea a seguir como trabajo futuro.

# 2 Calidad de la voz

## 2.1 Definición

La medición de la calidad consiste en evaluar cuan buena es una muestra de voz en una determinada tarea. La calidad es solo uno de muchos atributos de la señal de voz y suele confundirse frecuentemente con la inteligibilidad que es diferente, pues se refiere a lo que el locutor dice, o sea, el significado o contenido de las palabras [6].

Normalmente cuando una señal de audio llega al sistema auditivo humano se inicia un proceso de percepción, que solo puede ser descrito por el sujeto. El individuo establece una relación entre la calidad de lo que percibe y lo ideal, emitiendo un criterio. Por tanto la calidad de una muestra de voz es el resultado de un proceso de percepción y valoración en el ser humano [7]. Partiendo de esta idea la medición subjetiva de la calidad fue la primera variante aplicada para determinar la confiabilidad de una muestra, un grupo de expertos evalúa la muestra calificándola dentro de una escala predefinida, es lo que se conoce como medidas de calidad subjetivas. La calidad es por naturaleza altamente subjetiva, por lo que se hace complicado evaluar su confiabilidad. El motivo radica en que los oyentes encargados de evaluar la calidad pueden tener diferentes estándares de lo que se define como ¨bueno¨ o ¨malo¨,

dando lugar a una gran variabilidad en las calificaciones que asigne cada uno a la muestra de voz analizada [7]. Existe otro tipo de medidas de calidad: las medidas objetivas, que obtienen de forma automática un valor de calidad de la muestra de voz.

De la misma manera que el sistema auditivo humano, un sistema al realizar una comparación tienen en cuenta dos factores fundamentales que son su poder discriminativo y la cantidad de información de la que dispone. A su vez dentro de este último elemento se encuentran la cantidad de muestras adquiridas y la fidelidad o parecido con la muestra original.

En [5] se apoyan en un estándar de calidad para biometría que define la calidad según tres criterios. Uno de ellos es el carácter en el que se incluyen todas aquellas características físicas o del comportamiento del individuo que determinan la probabilidad de poder distinguirlo del otro, por ejemplo: distintividad, universalidad y actitud del individuo ante la captura de la muestra, entre otras. Otro elemento es la fiabilidad en la que se incluyen tres posibles procesos que degradan la calidad: adquisición de las muestras, procesado y extracción de las características. Por último la utilidad se basa en predecir qué influencia tendrá la calidad de una muestra de voz para el reconocimiento, partiendo del supuesto de que mientras mayor sea la calidad menor será la probabilidad de que el reconocedor cometa un error al clasificarla.

Es evidente cómo los factores que se relacionan con la fidelidad y el carácter tienen gran impacto en la calidad de una muestra de voz, pero la manera más eficaz de determinar el comportamiento de un método de medición de calidad, es evaluar su relación con el rendimiento del sistema de reconocimiento [5]. Por tal razón este trabajo se centra en la utilidad, evaluando la relación que existe entre la puntuación de reconocimiento y varias medidas de calidad a pesar de que todas no han sido probadas en sistemas automáticos de reconocimiento de locutores (SARL).

## 2.2 Factores que degradan la calidad de una muestra de voz

La principal causa de degradación de la calidad de la señal por los sistemas de comunicaciones modernos se debe esencialmente a la latencia, demora o retraso de la señal, pérdida de paquetes, variación de la demora *(jitter)*, eco y las distorsiones que se introducen por parte de la codificación [4]. Estos factores estarán presentes o no en dependencia de la aplicación en la que se quiere medir la calidad de una muestra. Dichos factores afectan parámetros de percepción subjetiva como es el caso de la inteligibilidad, naturalidad y volumen, que en su conjunto determinan la calidad de una muestra [7].

Un primer ejemplo es la manera en que se codifica la señal y la razón de codificación que se utiliza. Por ejemplo si se habla de la modulación diferencial adaptativa por impulsos codificados (*Adaptive Differential Pulse Code Modulation (ADPCM)*) su razón de bit es de 16 kbps mientras que la codificación basada en predicción lineal, por ejemplo la predicción lineal con excitación por código *(Code-excited linear predictio (CELP))* opera entre los 4 y 8 kbps. A medida que se reduce la razón de codificación de la señal, en búsqueda de un mejor aprovechamiento del ancho de banda, se afecta más la calidad de la muestra.

Puede darse el caso de que el efecto que provoca en una línea telefónica el sonido lateral o *sidetone*[1], pueda variar la confiabilidad de la muestra debido a que si existen pocas pérdidas de este sonido provoca que los niveles de señal que retornan sea demasiado altos e incómodos para el usuario, por el contrario si las pérdidas son excesivas puede que el teléfono suene muerto mientras se está utilizando.

Algunos algoritmos de compensación introducen distorsión en la señal como es el caso del ruido musical que introduce la sustracción espectral.

A continuación se muestran de manera más detallada otros elementos de gran importancia que atentan contra la calidad.

[1] El tono lateral de un aparato telefónico se basa en la transmisión de un sonido de fondo o adyacente desde un terminal hasta el otro.

**Factores relacionados con los sujetos (carácter)**
Como se comenta anteriormente las características físicas y de comportamiento de un individuo pueden afectar a la calidad de la muestra. Las que se relacionan con el sujeto se clasifican de la siguiente manera:

1. *Cambios en las características:* Pueden producirse cambios por distintos motivos: enfermedades o alteraciones emocionales. Por ejemplo una afonía o un constipado, que hacen cambiar las características de la voz.
2. *Comportamiento del sujeto:* Hay varias maneras por las que esto puede afectar a la calidad de una muestra: la cooperación del sujeto, su conocimiento sobre la captura de los parámetros y el estado emocional son ejemplo de ellas. En sistemas forenses de reconocimiento de locutor estos son factores especialmente importantes pues es común que el sujeto se encuentre en un estado emocional alterado, nervioso o que intente modificar su voz para evitar ser reconocido.
3. *Fraude:* Se pueden distinguir dos tipos:
    - Evasión: es muy común en casos forenses donde los locutores intentan distorsionar su voz de alguna manera para evitar ser identificados.
    - Suplantar la identidad (*Spoofing*): consiste en asumir la identidad de otra persona simulando su voz.

**Factores relacionados con la adquisición de datos**
Existen dos factores degradantes relacionados con la adquisición de datos:
1. *Dispositivos de adquisición de datos:* Distintos dispositivos de captura extraen diferente información con diversos grados de fidelidad. Como ejemplo evidente se encuentran las diferencias entre datos obtenidos de dispositivos microfónicos y telefónicos.
2. *Procesos de adquisición de datos:* Acá se encuentran factores tales como el tiempo de habla registrado o las condiciones ambientales de la adquisición (ambientes ruidosos, salas con reverberaciones, etc.) que influyen en la calidad de la muestra.

**Factores relacionados con el procesado y la compresión de los datos**
En los sistemas remotos de reconocimiento de locutores el procesamiento básico que recibe una señal de voz es su transmisión a través de una red telefónica que trae afectaciones en el ancho de banda, puede incorporar ruidos telefónicos, como es el caso de la diafonía o *crosstalk*. Por otro lado, a menudo se almacenan las muestras en formatos comprimidos, esta compresión puede dar lugar a pérdida de información, que puede conllevar a reducción de calidad en la muestra.

**Factores relacionados con la extracción de rasgos**
Diferentes métodos de extracción de rasgos arrojan diferente información sobre la muestra. Esta información vendrá acompañada por un grado de fiabilidad, un claro ejemplo se muestra al extraer una cantidad inapropiada de rasgos, que puede traer como consecuencia una baja calidad en la representación del locutor.

# 3 Clasificación de las medidas de calidad

## 3.1 Medidas subjetivas

Dentro de la literatura se han propuesto varios métodos subjetivos para medir la calidad. Estos pueden ser clasificados dentro de dos grupos: los que se basan en muestras de referencia y los que califican con un valor numérico la calidad de la muestra. En el primer caso se presenta a los oyentes un par de muestras, una de referencia y una de prueba. La muestra de referencia se construye normalmente a partir de filtrar o de añadir ruido a una muestra limpia. Los expertos tendrán que seleccionar la que ellos consideren con mejor calidad.

En el caso de las calificaciones numéricas se le presenta a los sujetos una muestra de prueba, y se pide que califiquen la calidad de esta, típicamente en un rango entre 1 y 5, donde 1 corresponda a la más baja calidad y 5 a excelente. No se necesita una muestra de referencia en este caso. Ambas maneras de medir la calidad tienen sus ventajas y limitantes, en la práctica su uso estará en dependencia de la aplicación.

### *3.1.1 Puntuación de opinión media*

Este es el método que más se utiliza para determinar la calidad de manera subjetiva y se incluye dentro del grupo que califica con un valor numérico la calidad de la muestra. Se asignan valores en una escala entre 1 y 5, según la distorsión perceptible en la muestra (ver tabla 1). Es uno de los métodos recomendados por IEEE[1] y la UIT [3]. La medida de calidad obtenida es el resultado de promediar la puntuación asignada por cada oyente. Este valor se conoce como MOS *(Mean Opinion Score (MOS))*. La prueba consta de dos fases, entrenamiento y evaluación. En la primera fase los expertos escuchan un conjunto de muestras de referencia que ejemplifican los casos de excelente, media y baja calidad. Esta etapa es de alta importancia para poder ajustar el rango en que calificará cada sujeto que participa en la prueba. En la fase de evaluación se califican las señales de prueba dentro de las categorías que se muestran en la Tabla 1.

**Tabla 1.** Escala MOS.

| Índice | Calidad de la muestra | Nivel de distorsión |
|---|---|---|
| 5 | Excelente | Imperceptible |
| 4 | Buena | Sólo perceptible, pero no molesta |
| 3 | Razonable | Perceptible y un poco molesto |
| 2 | Pobre | Molesto pero no desagradable |
| 1 | Malo | Muy molesto y desagradable |

Los expertos se ven forzados entonces a incluir la calidad de cada muestra en una de estas 5 categorías, asumiendo que están uniformemente espaciadas. Esta asunción sin embargo puede no cumplirse por lo que se han hecho modificaciones para obtener una mayor exactitud [2].

## 3.2 Medidas objetivas

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 de voz sobre IP, o a un sistema de reconocimiento de locutores donde no se cuenta con la seña original. 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. En la Tabla 2 se resumen las ventajas y limitantes de ambos tipos de medidas.

**Tabla 2.** Comparación entre medidas de calidad objetivas y subjetivas. El símbolo ¨+¨ denota la ventaja de un método sobre otro, señalado con ¨-¨.

| | **Medidas subjetivas** | **Medidas objetivas** |
|---|---|---|
| Costo | - | + |
| Reproducibilidad | - | + |
| Automatización | - | + |
| Alteraciones imprevistas | + | - |

Idealmente las medidas de calidad objetivas deben ser capaces de calificar la muestra sin necesidad de acceder a la señal original, además de hacerlo con la misma exactitud que lo hacen las pruebas subjetivas. Estas medidas deben incorporar información de alto y bajo nivel. Los bajos niveles de información reflejan las características acústicas del habla por lo que sería recomendable que las medidas de calidad incluyan información acústica. Los altos niveles de información reflejan los hábitos y el estilo del locutor al hablar los que son mucho más complejos y difíciles de describir. Para poder determinarlos en una muestra se requieren largos intervalos de tiempo, siendo más robustos a diversas distorsiones, por ejemplo al ruido, es por esta razón que es útil que las medidas de calidad incluyan además información prosódica, fonética, semántica y pragmática, pues estarían modelando parámetros mucho más discriminativos del locutor. Además los seres humanos reconocen a un locutor mezclando ambos tipos de información de la muestra de voz, por lo que sería muy conveniente poder mezclar ambos tipos de información en las medidas de calidad [7, 8]. Por este motivo las medidas de calidad objetivas que se diseñen deben estar altamente correlacionadas el resultado de aplicar una medida subjetiva para determinar la calidad. Una manera de corroborar esto es a partir de bases de datos etiquetadas con dichos valores de calidad, por lo general en escala MOS [9].

Las medidas objetivas se clasifican en intrusivas o no intrusivas en dependencia de requerir o no una muestra de voz original, sin ser procesada [10]. 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. Tienen entre sus limitantes la necesidad de requerir la señal original, y el hecho de que además solo modelan bajos niveles de información. Las medidas de calidad objetivas no intrusivas tratan de establecer la calidad de una muestra procesada, utilizando métodos que evalúan la misma sin tener en cuenta la muestra original. 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 [11, 12] se han encaminado a desarrollar medidas objetivas que modelen varios aspectos del sistema auditivo del hombre, que es el ejecutor de la medida subjetiva. La figura 1 muestra de qué manera se aplica ambos tipos de medidas.

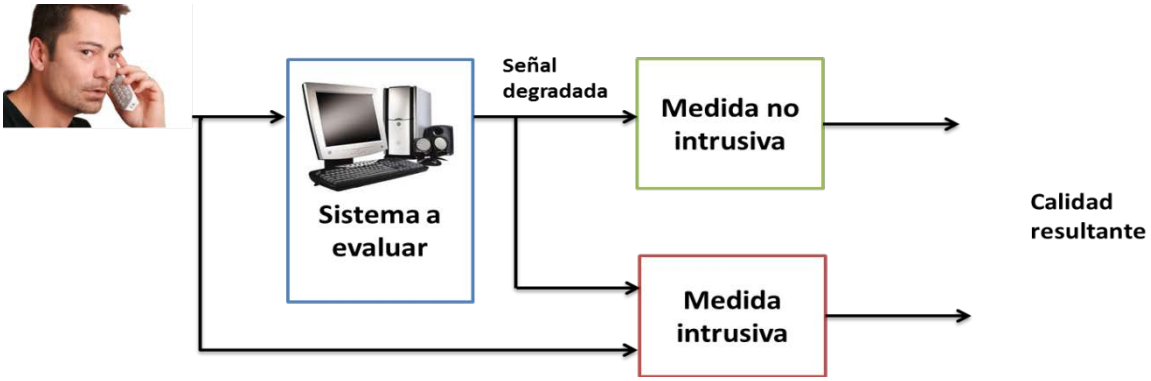


**Fig. 1.** Esquema de aplicación de medidas de calidad intrusiva y no intrusiva.

Cada medida, intrusiva o no, devuelve valores en diferentes rangos, por lo que es necesario ajustarlas a un intervalo común para poder analizar sus resultados. Normalmente 0 corresponderá al peor caso y 1 a la máxima calidad. Para hacer este ajuste se diseña una función de mapeo $Q(x)$ para cada medida de distorsión, donde $x$ corresponde a cada posible valor de degradación que se obtenga de aplicar la medida.

### *3.2.1 Medidas de calidad objetivas intrusivas*

La aplicación de medidas de calidad intrusivas estará en dependencia de la aplicación en la que se desee determinar la calidad de la muestra.

Las medidas de distorsión intrusivas, como también se le conocen, pueden aplicarse tanto en el dominio del tiempo como en el de la frecuencia. En el dominio de la frecuencia se asume que cualquier distorsión o diferencia detectada en el espectro de magnitud esta correlacionada con la voz. Dentro de las medidas de calidad intrusivas resaltan las siguientes:

1. Medidas basadas en la relación señal ruido *(Signal to Noise Ratio (SNR))*
- SNR segmental en el dominio del tiempo [7].
- SNR segmental en el dominio de la frecuencia [4].

2. Medidas basadas en la distancia espectral a partir de los coeficientes de predicción lineal *(Linear Prediction Coefficients (LPC))*
- Razón logarítmica de similaridad *(Log-likelihood ratio(LLR))* [13].
- Distancia de Itakura-Saito *(Itakura-Saito Distance(IS))* [14].
- Distorsión espectral *(Spectral Distorsion(SD))* [13].
- Distancia Cepstral *(Cepstral Distance(CD))* [13].

3. Medidas basadas en la percepción
- Medidas de distorsión de Bark *(Bark distorsion measures (BSD))* [11].
- Evaluación perceptual de calidad del habla *(Perceptual Evaluation of Speech Quality(PESQ))* [12].

4. Medidas compuestas [7].

La mayor parte de las medidas se aplican sobre segmentos de la señal de voz que oscilan entre los 10 y 30 ms, (bajo nivel de procesamiento) comparando por tramas la calidad de la señal procesada respecto a la señal original. Luego se obtiene un solo valor de calidad promediando los valores obtenidos por cada trama.

La SNR puede evaluarse tanto en el tiempo como en la frecuencia, en el tiempo es una de las medidas objetivas más simples y usadas para evaluar métodos de compensación o de codificación de la

voz. Para que esta medida sea lo suficientemente efectiva es necesario que la señal original y la procesada estén alineadas en el tiempo y cualquier error que exista en la fase se corrija. Uno de los principales problemas de esta medida está dado por el hecho de que la energía de la señal en los intervalos de silencio será pequeña por lo que pudieran obtenerse valores de SNR negativos. Estos intervalos de silencio en señales de voz son muy comunes por lo que los valores negativos de SNR que se obtengan de las tramas de silencio influirán en el resultado final de la medida, pues no estaría reflejando la calidad de la información que realmente se utiliza para reconocer al locutor. Una manera de solucionar este problema puede ser eliminando las tramas de silencio de la comparación que se hace en la medida de distorsión, utilizando métodos de detección de actividad de voz *(Voice Activity Detection (VAD)).*

Las medidas en el dominio de la frecuencia presentan una mejor correlación con la percepción de la voz por el sistema auditivo, siendo fáciles de implementar. Una de sus principales ventajas está dada porque son menos sensibles a la no alineación de las señales en el tiempo. En este grupo se incluyen además de la medida de SNR segmental, la LLR, la distancia de IS, la SD y la CD. Es necesario señalar que todas estas medidas solo son válidas cuando se aplican a las tramas de voz de la señal, debiendo eliminarse previamente los intervalos de silencios con el uso del VAD. Las medidas anteriores son simples de implementar y fáciles de evaluar, sin embargo su habilidad de semejarse a la medida de calidad subjetiva, por ejemplo, tratando de ajustarlas a la escala MOS, es limitada. Esto se debe a que no evalúan el procesamiento de la señal que ocurre en la periferia del sistema auditivo. Entre los elementos que no modelan están la selectividad de frecuencias propia de la frecuencia normal de escucha y el ruido que se percibe. Por tal razón existen otras medidas objetivas que modelan la percepción de habla y su calidad como es el caso de la BSD y la PESQ. Estas medidas tienen en cuenta las propiedades más importantes del procesamiento auditivo del habla como la resolución no uniforme de la frecuencia de percepción, esta normalmente se modela a través de un banco de filtros pasa-banda cuya frecuencia central y ancho de banda varían de acuerdo con la frecuencia en escalas Bark o Mel.

La medida BSD es un promedio de la distancia Euclidiana entre la señal original y la distorsionada en el dominio Bark.

$$BSD(k) = \sum_{b=1}^{N_b} \left[ S_k(b) - \overline{S_k(b)} \right]^2 \quad , \tag{1}$$

donde $S_k(b)$ $y$ $\overline{S_k(b)}$ son el espectro sonoro de la señal limpia y de la corrupta respectivamente y $N_b$ es el número de bandas críticas. Esta medida produce altos valores para los segmentos no sonoros de la voz por lo que se excluyen del análisis utilizando un VAD. Es de destacar que esta medida tiene una alta correlación ($\rho > 0.9$) con la escala MOS.

La medida PESQ es propuesta por la UIT-T en la recomendación P.862 [12] en sustitución a otro grupo de recomendaciones que tenían un alcance muy limitado. Esta surge con la idea de tener en cuenta las distorsiones que se producen en la señal cuando esta viaja a través de una red de telecomunicaciones, como es el caso de la perdida de paquetes, las demoras y la codificación. Tienen una alta correlación ($\rho > 0.92$) con las pruebas subjetivas de escucha. Sin embargo no evalúa totalmente la calidad de la señal al trasmitirse a través de una red telefónica porque solo toma en cuenta la señal cuando viaja en un solo sentido. Efectos tales como la pérdida de la sonoridad, el mal posicionamiento del tono lateral y el eco no se tienen en consideración.

Por último deben mencionarse las medidas compuestas. Se basan en combinar múltiples medidas objetivas, pues si diferentes medidas abarcan diferentes características de la señal distorsionada, combinarlas trae significativas mejoras en su correlación con las pruebas subjetivas.

A pesar de que las medidas compuestas siempre mejoran la correlación con las medidas subjetivas, es necesario usarlas con un determinado conjunto de muestras distorsionadas solo para la prueba y

diferentes a las que fueron usadas para entrenar. Esto se debe a que si se realiza la prueba bajo las mismas condiciones que el entrenamiento se estaría condicionando el resultado.

### *3.2.2 Medidas de calidad objetivas no intrusivas*

Las medidas objetivas mencionadas anteriormente son intrusivas por naturaleza ya que requieren la señal original para realizar la comparación. Estas predicen la calidad estimando la distorsión que existe entre la señal de entrada (limpia) y procesada (corrupta) para luego mapear el resultado estimado en una métrica de calidad.

Sin embargo, en algunas aplicaciones la señal original no está realmente disponible por lo que las medidas objetivas intrusivas no son útiles. Por ejemplo si se desea determinar la calidad en aplicaciones 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 sistemas automáticos de reconocimiento de locutores 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 de reconocimiento de locutores como son el control de acceso, la autenticación del usuario o el manejo personalizado de datos, 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 la original es bien diferente y han surgido varias medidas de distorsión objetivas no intrusivas. Algunas de ellas modelan el tracto vocal para evaluar la distorsión, unas parten de la manera en que se codifica la señal mientras que otras evalúan un conjunto de ruidos o distorsiones que se presentan en la señal, para emitir un criterio sobre su calidad.

A continuación se presentan las principales medidas objetivas no intrusivas

1. Según se estima el ruido
    - SNR(determinada a partir de los silencios) [15]
    - SNR-Wiener(a partir de un filtro adaptativo) [10]
    - Método del Histograma de Hirsch [16]
    - Distancia entre armónicos *(Harmonicity Distance(HD))* [15]
2. Medidas estadísticas
    - Kurtosis global [17]
    - *Skewness* global [17]
    - Kurtosis a partir de histograma (Kbin) [17]
    - Kurtosis local [18]
    - *Skewness* local [18]
    - Kurtosis sobre los LPC [19]
    - *Skewness* sobre LPC [18]
    - Kurtosis sobre los coeficientes Cepstrales [18]
    - *Skewness* sobre los coeficientes Cepstrales [18]
3. Medidas basadas en modelos estadísticos
    - UBML (Universal Background Model Likelihood) [5]
    - Medida de calidad de mejora no intrusiva usando modelos de degradación [20]
4. Medidas basadas en el análisis del tracto vocal
    - Medida no intrusiva para determinar la calidad de la voz basada en modelos del tracto vocal [21]

5. Medidas de estimación de la calidad subjetiva
    - UIT-T P.56 [18]

La finalidad de las medidas que se agrupan de acuerdo a la manera en que se estima el ruido, es determinar el nivel que este tiene en una muestra de voz. La mayoría de las medidas de calidad están ligadas a este nivel y la más común es la SNR. Para calcularla existen diversas maneras, como es determinar la energía del ruido en los silencios de la señal y la energía de voz en las zonas de habla, para después calcular la energía media en cada uno de ellos. Sin embargo a pesar de ser sencillo puede darse el caso de que las muestras tenga alta actividad vocal por lo que no tendrá suficiente tiempo de silencio para estimar adecuadamente el nivel de ruido. Además no se tienen en cuenta ruidos que existen en las regiones donde está presente la voz, que pueden ser causados por micrófonos, amplificadores, codificadores, etc.

Las medidas estadísticas consisten en obtener diferentes tipos de evaluación estadística, ya sea en la voz o en parámetros de esta, que sea indicativa de la degradación de la señal. En [18] se utilizan el Skewness y la Kurtosis las cuales se aplican a los coeficientes cepstrales en escala Mel(Mel Frequency Cepstral Coefficients(MFCC)) y a los LPC de la señal. Además se aplican estas mismas medidas estadísticas en el dominio temporal con buenos resultados [17].

Las medidas basadas en modelos estadísticos aprovechan los modelos de habla poco degradada para determinar la calidad. Es el caso del método propuesto en [5] se trata de aproximar 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. 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. En el caso de [20] se propone un algoritmo más exacto y robusto para determinar la calidad basado en GMM. La robustez y exactitud se logran equipando el algoritmo con información relativa a cuan corrupta puede estar la señal debido a diferentes esquemas de transmisión y codificación, así como información relativa a señal no corrupta.

En las estimaciones basadas en el análisis del tracto vocal se pretende identificar las distorsiones presentes en las muestras basándose en modelos del tracto vocal asociados a dichas muestras. Estos modelos se realizan asociando cada cavidad del tracto con un tubo de diámetro X y longitud Y que varía con el tiempo. Para esto se definen una serie de reglas donde se identifican las violaciones de lo que se considera habla normal. Partiendo de que se dispone de una representación del tracto vocal como la que ya mencionada anteriormente, una posible violación estará dada por un incremento significativo del diámetro de una cavidad en un pequeño espacio de tiempo. Este método se ha encontrado en [18, 21], el primero corresponde con la recomendación P.563 de la UIT-T en la que se extraen 14 parámetros para estimar la calidad de la señal. El segundo modela el tracto vocal de una manera diferente lo que permite predecir la calidad de la señal degradada con gran exactitud.

El objetivo de los métodos que estiman la calidad subjetiva es estimar la calidad de la señal aproximándose lo más posible al modelo psicoacústico humano. Suelen ser medidas enfocadas a determinar la calidad en redes de telecomunicaciones. En este grupo se encuentra la P.563 de la UIT-T que ha sido utilizada en sistemas de reconocimiento de locutores [5, 8].

# 4 Medidas de calidad utilizadas en reconocimiento de locutores

Las medidas de calidad objetivas no intrusivas se han diseñado para el monitoreo de la calidad en tiempo real en una red de telecomunicaciones y no para ser usadas en el reconocimiento de locutores. A pesar de esto, varios estudios realizados han tomado algunas de ellas y las han aplicado al reconocimiento de locutores. Estas se detallan a continuación.

## 4.1 Kurtosis LPC (KLPC)

La Kurtosis o momento estadístico de cuarto orden es una medida de la elevación o achatamiento de la distribución que siguen los valores reales de una variable aleatoria. A mayores valores, mayor será el pico de la distribución por lo que existe una mayor concentración de los datos muy cerca de la media de la distribución [22]. El valor de la kurtosis de una distribución gaussiana es 3, esta se toma como referencia para clasificar otras, permitiendo medir básicamente el parecido entre una distribución cualquiera y una gaussiana. En este caso la kurtosis se aplica a la distribución de los coeficientes LPC en cada trama de la señal, de la misma manera que se hace en la UIT-T P.563 [18] y en [19].

Para cada trama de 20 ms de la señal se obtienen P coeficientes $a_p$, luego la kurtosis por trama se determina de la siguiente manera:

$$k = \frac{1}{P}\sum_{p=1}^{P}\left(\frac{a_p - \frac{1}{P}\sum_{p=1}^{P} a_p}{\sigma}\right)^4 \quad , \tag{2}$$

donde $\sigma$ representa la desviación estándar de los coeficientes en la trama, por último se promedian todos los valores de Kurtosis obtenidos para cada trama de voz en la señal. Como se comprobará posteriormente en los resultados experimentales, la eficacia del reconocimiento del locutor disminuye a medida que el valor de KLPC aumenta. O sea, a medida que la distribución de los coeficientes LPC por trama de la señal, sea más gaussiana (más baja y cercana a 3), mejor calidad presentará la señal para ser utilizada en el reconocimiento de locutores.

La función de mapeo $Q_{KLPC}(x)$ es la siguiente:

$$Q_{KLPC}(x) = 1 - \frac{x - min}{max - min} \ , \tag{3}$$

donde $max$ y $min$ corresponden con los valores máximos y mínimos alcanzados por la medida.

## 4.2 Kurtosis Cepstral (KCEP)

Esta medida de muy semejante a la anterior solo que ahora la kurtosis se aplica sobre los coeficientes MFCC para cada trama de voz de la señal que se analiza [5, 18, 23]. Tiene además la misma interpretación que la kurtosis sobre los LPC por lo que la función de mapeo $Q_{KCEP}(x)$ es la misma en este caso.

## 4.3 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) y se puede definir 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. Este algoritmo es aplicable para predecir la calidad sin una señal de referencia, por este motivo se recomienda para la evaluación no intrusiva de la calidad para la supervisión y evaluación de una red en funcionamiento, empleando en el extremo lejano de una

conexión telefónica, fuentes de señal desconocidas. Hasta el momento en que surge esta recomendación los métodos de evaluación de la calidad de la voz para sistema telefónicos como la Recomendación UIT-T P.862 requerían una señal de referencia o bien solo calculaban el índice de calidad basado en un conjunto restringido de parámetros tales como el nivel de ruido, ruido en las pausas de la voz y el eco.
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 MOS de acuerdo con la Recomendación UIT-T P.800.1. No está limitada a aplicarse exactamente a los extremos, puede utilizarse en cualquier lugar de la red, por tanto la puntuación calculada es comparable con la calidad percibida por un oyente humano que escucha en ese punto con un auricular convencional.

En la Recomendación se definen las condiciones en las que puede aplicarse y sus limitantes. Entre los señalamientos más importantes se destaca que a pesar de que el valor de correlación entre la medida objetiva y la subjetiva sea de 0.89, el algoritmo no debe utilizarse para sustituir pruebas subjetivas, pero puede aplicarse para realizar medidas cuando las pruebas de audición resulten excesivamente caras o no aplicables en lo absoluto. Dado que P.563 modela la percepción de la calidad observada por el ser humano empleando un terminal receptor común, no puede tenerse en cuenta la degradación que pueda introducir el terminal de recepción ni cualquier otro equipo en una conexión real objeto de supervisión.

Por otro lado, los factores que degradan la calidad se dividen en tres grupos, los que dificultan la posibilidad de escuchar, de hablar y de establecer una conversación. En el primer grupo se incluyen: perdida de paquetes, distorsión producida por la codificación recortes de la señal y eco del que escucha. En el segundo se pueden encontrar el eco del hablante y el mal posicionamiento del tono lateral, mientras que en el tercero están incluidos la demora, lo que se conoce como el cruce de las líneas y el ruido de fondo. Puesto que P.563 predice las notas de calidad de audición, no pueden tenerse en cuenta los efectos que solo degraden la calidad del habla o de la conversación. Ello significa que los efectos de la pérdida de sonoridad, retardo, tonos laterales, eco del hablante y cualesquiera degradaciones que afecten exclusivamente a la calidad vocal o a las interacciones bidireccionales, no se reflejan en las puntuaciones de P.563. Por esta razón es probable que aunque se obtengan elevadas puntuaciones, la calidad global de la conexión no sea óptima.

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 existen condiciones para los que esta recomendación devuelve resultados inexactos, por ejemplo el efecto del retardo en conversaciones, y 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.

Este algoritmo fue solamente diseñado para procesar la voz humana, no es capaz de evaluar ruido, música o cualquier señal de audio no vocal, pero pueden utilizarse los resultados al simular la transmisión de voz u otros procesamientos siempre que se encuentren dentro de las condiciones de validación mencionadas anteriormente. La señal digitalizada debe tener frecuencia de muestreo de 8000 Hz, 16 bit/muestras y su duración debe oscilar entre los 3 y los 20 segundos.

La señal se procesa de varias maneras, a modo de capas, que detectan un grupo de parámetros de señal característicos. 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 2 resume el proceso por el que transita la señal hasta obtener un valor final de calidad en la escala MOS.

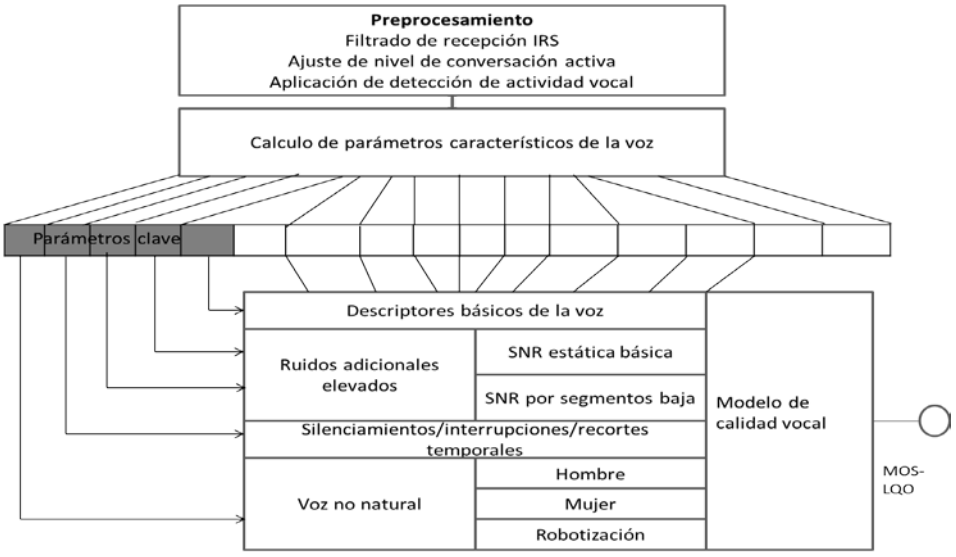


**Fig. 2.** Diagrama en bloques de UIT-T P.563

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.

*4.3.1 Análisis del tracto vocal y desnaturalización de la voz.*

- Este bloque trata de detectar el carácter desnaturalizado de la voz, contiene un modelo de producción vocal para extraer partes de la señal que podrían interpretarse como voz y separarlas de las partes no vocales. Además, el análisis estadístico de orden superior ofrece información adicional sobre el grado de humanización de la voz analizada. El carácter desnaturalizado de la voz se valora de forma separada para voces masculinas y femeninas. Además, en caso de intensa robotización, se realiza otra valoración adicional, independiente del género.

- Se analiza la señal para detectar la presencia de tonos tales como los tonos DTMF (dual-tone multi-frequency signaling) 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. Algunos códecs vocales utilizan métodos de “compensación” de errores sustituyendo un paquete perdido por un paquete anteriormente transmitido con éxito, lo cual tiende a disminuir la calidad de la señal en lugar de aumentarla.

*4.3.2 Análisis del ruido adicional intenso*

- El análisis del ruido calcula distintas características del mismo. En base a dos parámetros fundamentales se decide si el ruido adicional es la degradación principal. Si se detecta que el ruido adicional es la principal causa de degradación, se toma una decisión acerca del tipo de ruido. 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 detectara la presencia de ruido con un carácter probablemente estático, existen varios detectores para cuantificar la cantidad de ruido a nivel 'local' y 'global'. La expresión de ruido 'local', describe al ruido que se encuentra fundamentalmente en los fonemas, mientras que ruido 'global' se define como el ruido existente entre conjuntos de sonidos tales como frases. La distinción entre ambos tipos de ruido es importante, ya que en comunicaciones móviles a menudo se aplican criterios diferentes para las partes activas y no activas de la voz, por ejemplo, introduciendo ruido confortativo.

#### *4.3.3 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 en la señal.

- 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. El recorte es un fenómeno molesto que elimina un bit de la señal de voz en el instante en que el transmisor detecta la presencia de señal vocal. Es posible detectar las interrupciones de la señal vocal producidas durante los intervalos de señal vocal activa.

Cada uno de los bloques funcionales descritos anteriormente se cuantifica utilizando una serie de parámetros de la voz. Por ejemplo, las estadísticas de segundo, tercer y cuarto orden medidas sobre los coeficientes LPC del tracto vocal, diferentes medidas de energía de la señal y otros que aparecen listados en la recomendación [18]. Incluso para el preprocesamiento y el VAD se calculan varios de estos parámetros. Algunos de ellos son medidos en determinado bloque funcional y luego se reutilizan en uno que le sucede.

La función de mapeo se define de acuerdo con la escala MOS:

$$Q_{P563}(x) = \frac{(x-1)}{4} \quad . \tag{4}$$

### 4.4 Distancia entre armónicos (Harmonicity Distance HD)

La estructura de los armónicos puede ser fácilmente reconocida en una representación espectrográfica de la voz, incluso en entornos ruidosos. La armonicidad denota cuan armónica o periódica es la señal además de cuan sonora es. Cuando la voz esta corrupta por ruido aditivo [24] la estructura de los armónicos se afecta en dependencia del tipo de ruido y nivel, determinando el grado y distribución de la distorsión. Cabe destacar entonces que a medida que disminuye la armonicidad en la señal mayor es la distorsión y menor la calidad [25].

Para medir la distorsión en la estructura de los armónicos de la voz se define una función que relaciona la potencia de los armónicos y la potencia entre ellos. Esta función se define de la siguiente manera:

$$HD = \frac{1}{NH \times N_{frames}} \sum_{N_{frames}} \sum_{k=1}^{NH} 10log \frac{P_k + P_{k+1}}{2P_{k,k+1}} \quad , \tag{5}$$

donde $P_k$ es la potencia en el armónico *k*, $P_{k,k+1}$ es la potencia entre los armónicos *k* y *k+1*, *NH* representa la cantidad de armónicos y $N_{frames}$ es la cantidad de tramas de voz en la señal. La función de mapeo se define de la siguiente manera:

$$Q_{HD}(x) = \frac{x-min}{max} \tag{6}$$

El uso de esta medida no se encuentra reportado en ningún trabajo, se ha seleccionado por ser una medida de SNR en la señal, debido a las características que evalúa y el lugar donde lo hace.

# 5 Relación de las medidas de calidad con el reconocimiento de locutores

Las medidas de calidad que se relacionan en los epígrafes anteriores no han sido diseñadas para el reconocimiento de locutores, sin embargo varios autores han vinculado algunas de dichas medidas a los resultados de este tipo sistema desde distintos enfoques. En [26] se relacionan el envejecimiento, la calidad y el resultado la de verificación con el objetivo de observar la influencia que tienen, tanto juntos como por separados, estos indicadores en el reconocimiento de locutores. Para ello se identifican parámetros que varían con la edad como son la disminución de la frecuencia fundamental y cambios en el timbre, demostrando que la edad es un factor de variabilidad en el resultado, pero que si se combina con una evaluación de la calidad los resultados serán significativamente mejores.

Las medidas capaces de representar de manera más fiel las características discriminativas del locutor así como las que tienen una relación más directa con el comportamiento de un sistema de reconocimiento de locutores han sido las seleccionadas para aplicarlas a los mismos.

## 5.1 Medidas de calidad aplicadas durante el cálculo y la fusión de *scores*. Inclusión de diferentes niveles de información en el reconocimiento del locutor, en dependencia de su calidad.

Existe una clara relación entre el sistema de reconocimiento y los diversos niveles de información en la señal de voz debido a que los seres humanos parten de varios tipos de información para reconocer la identidad de un locutor. A partir de esta idea y de la estructura de un sistema de reconocimiento es posible incluir la información relativa a la calidad en las etapas de extracción de rasgos, entrenamiento de los modelos, determinación de la puntuación y fusión de estas [27].

Trabajos previos [28, 29] han mostrado resultados alentadores cuando se incorpora la calidad en el proceso de reconocimiento sobre todo en las dos últimas etapas del sistema. Sin embargo en [8] ya se incluye información referente a la calidad de las muestras en la fusión de los scores y de diferentes niveles de información así como en agregar esta información al proceso de determinación de la puntuación. Usualmente para el cálculo de la puntuación se utiliza un sistema GMM-UBM, utilizando una etapa de pre-procesamiento en las que se ejecutan dos tareas principales: mejora de la señal a partir de eliminar los efectos del canal y de reducir el ruido en la muestra y eliminación de los silencios y los sonidos que no se consideran voz, preservando solo la información que satisfaga un determinado criterio y eliminando el resto. Si se combina esta etapa con un mecanismo clásico para determinar la puntuación, se le confiere a toda la información que se preserva, luego de la etapa de preprocesamiento, 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 [30]. Si la puntuación se calcula basado en la calidad (Quality-Based Score Computation (QBSC) [8] 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, la figura 3 muestra este vínculo, aplicando las medidas de calidad en la etapa de preprocesamiento.

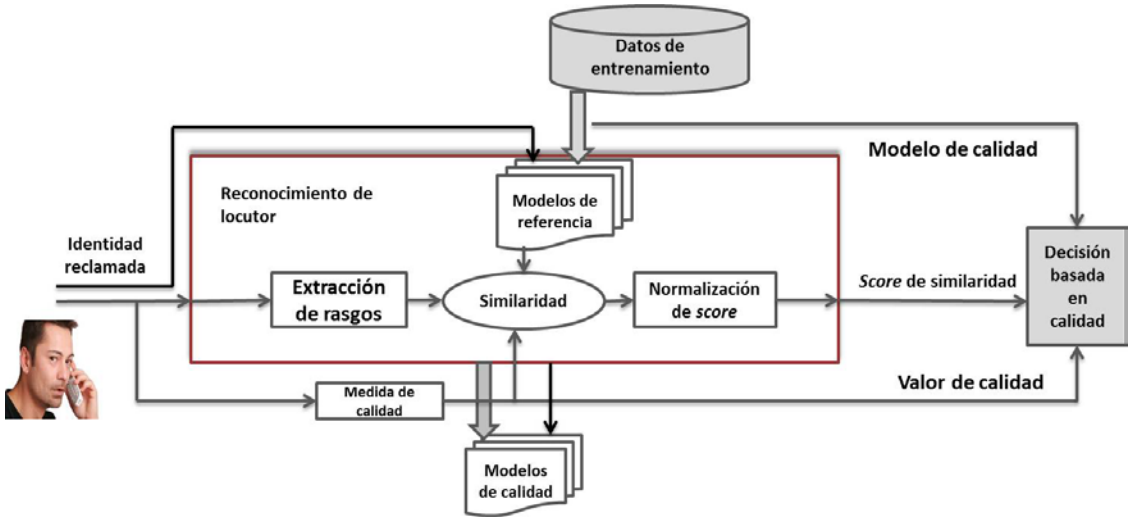


**Fig. 3.** Modelo general de un sistema de reconocimiento de locutor utilizando medidas de calidad.

Dado una secuencia de vectores de rasgos $O = \{o_1, o_2, o_3, \dots o_T\}$ y su correspondiente vector de calidad, $Q^{\xi} = \left\{q_1^{\xi}, q_2^{\xi}, \dots . q_T^{\xi}\right\}$ calculada para cada trama de la señal utilizando la medida de distorsión $\xi$. La probabilidad de la secuencia de vectores de rasgos respecto al modelo $\lambda$ incorporando la medida de calidad como un factor de peso se determina de la siguiente manera:

$$p(O|Q,\lambda) = \prod_{t=1}^{T} p(o_t|\lambda)^{q^{\xi}} \quad . \tag{7}$$

Luego el logaritmo de la probabilidad se determina como:

$$\log p(O|Q,\lambda) = \sum_{t=1}^{T} q_t^{\xi} \log p(o_t|\lambda) \quad . \tag{8}$$

Si se trata de fusión de scores combinando varios niveles de información se utiliza una Máquina de Soporte Vectorial (Support Vector Machine (SVM)) adaptada para poder incluir la información relativa a la calidad a partir de varios niveles de información. Para esto el método se basa en la combinación de información de bajo nivel del locutor (información espectral, por ejemplo) con otros tipos de información de alto nivel (información fonética y lexical, por ejemplo). 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 caso, por lo que el diseño de las medidas de calidad sea más sencillo para los sistemas que utilizan bajos niveles de información.

A partir de esta idea se propone el modelo de la figura 4 llamado fusión de scores basado en la calidad (Quality-Based Score Fusion (QBSF)) [8] 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.

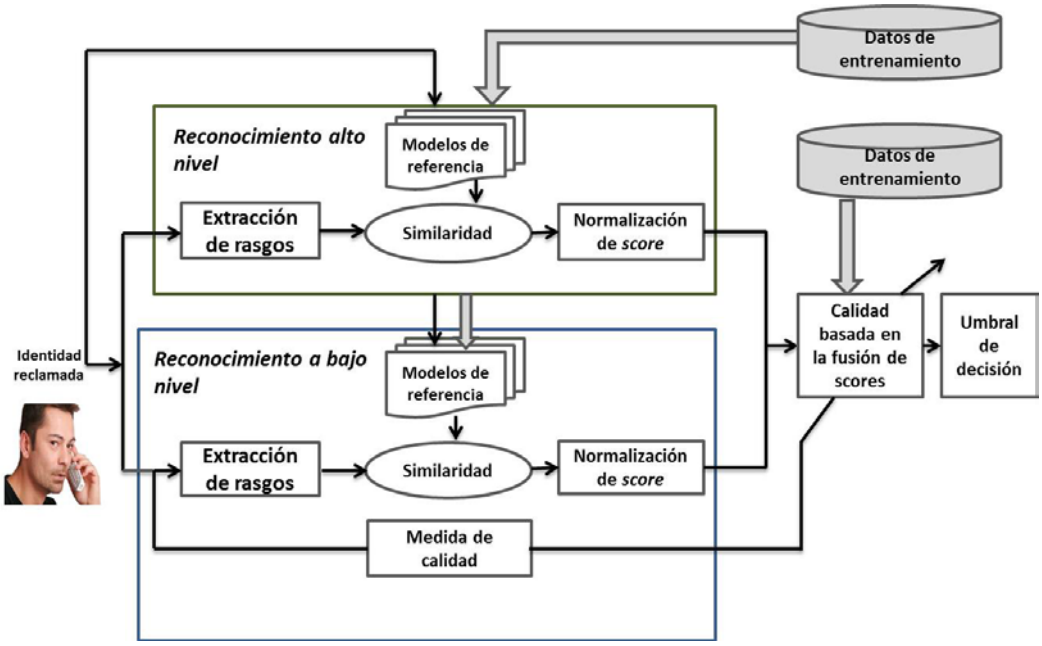


**Fig. 4.** Modelo general de un sistema de reconocimiento de locutores basado en varios niveles de información

Existen otros trabajos [29] en los que cualquier rasgo o característica que tenga baja calidad es eliminado completamente. En este caso la información de bajo nivel nunca se descarta y se debe a que este tipo de información brinda mejores resultados que cualquier otro sistema. En este sentido el valor de calidad determina si la puntuación final se calcula solamente basada en información de bajo nivel o si se realiza una combinación de ambos. Esta modificación implica que la puntuación resultante será al menos tan exacto como el del sistema de mejor desempeño, o mejor.

## 5.2 Métodos para verificar la confiabilidad en la decisión de un sistema de reconocimiento de locutor a partir de medidas de calidad

El objetivo de este método es estimar la confiabilidad de la decisión de un sistema de verificación en cada comparación que se realiza. Para esto se descartan las pruebas no confiables para asegurar que la decisión tomada con cada comparación implique el más bajo error. La confiabilidad de la comparación se obtiene a partir de medidas de calidad para la fase de entrenamiento y prueba.

Una manera común de determinar la relación entre la puntuación de verificación y la medida de calidad es a través de una red bayesiana donde intervienen los valores de dichos parámetros partiendo de la idea de que los factores que degradan la señal afectan de manera diferenciada a los clientes y a los impostores del sistema [31]. Existe otra propuesta que estima la confiabilidad de la prueba partiendo de una red bayesiana pero utilizando el resultado obtenido de combinar varias medidas de calidad [32].

### 5.3 Métodos para compensar la disminución en el rendimiento de un sistema de reconocimiento de locutores a partir de medidas de calidad

La idea de compensar el rendimiento de un sistema de reconocimiento a partir de los *scores* de verificación se debe a la gran variabilidad en la calidad de las muestras de audio comparadas, que puede ser introducida por diversos factores. Existen técnicas para compensar la variabilidad introducida por el canal como es el caso del Análisis de Factores [33] que depende en su mayoría de la selección de un corpus apropiado, preferiblemente con las mismas condiciones de la voz a reconocer.
Sin embargo la calidad de la voz permite predecir tanto el rendimiento de un sistema de verificación de locutor como una posible desalineación de las puntuaciones del mismo debido a cambios en dicha calidad. Las puntuaciones de los clientes e impostores pueden desajustarse y la calidad es una variante para realizar un ajuste de esta diferencia.

En [34] se propone utilizar modelos de regresión logística para este fin. En el método propuesto se utilizan varias medidas de calidad entre las que está incluida la SNR, una vez que se evalúa el comportamiento del sistema con varios indicadores de calidad, se proponen tres algoritmos basados en regresión logística: Regresión logística lineal de dos dimensiones (2D-LLR) y la Regresión logística bilineal (BLR tipo 1 y tipo 2) para evaluar el rendimiento del sistema. En esta propuesta se realiza una regresión logística para subconjuntos de *scores* en dependencia de la calidad del modelo del locutor entrenado y del locutor a verificar $Q_{train}$ y $Q_{test}$.

Trabajos más recientes [5, 19] también manejan estos términos, definiendo la calidad de una comparación, o calidad de la puntuación como:

$$Q = \sqrt{Q_{train} Q_{test}} \quad . \tag{9}$$

A partir de esta definición se desarrolla una metodología para determinar la relación entre estos parámetros y la puntuación, analizando el comportamiento de subconjuntos de *scores* agrupados teniendo en cuenta el valor medio obtenido para cada medida de calidad utilizada en el análisis. Obteniendo finalmente un $EER_K$ y una media $\mu_k$ para cada subconjunto, por cada medida de calidad utilizada, con el objetivo de comparar gráficamente el comportamiento para diferentes medidas de calidad. Este método permite analizar si una medida es representativa de la degradación de las muestras de voz, esto es posible si se determina una relación entre dicha medida y el rendimiento del sistema, lo que se conoce como impacto y se define como:

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

donde $EER_{max}$ se refiere al mayor valor de score obtenido para una medida de calidad determinada mientras que $EER_{min}$ 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.

Se realiza además una evaluación de la correlación entre las medidas de calidad con el fin de observar cuan complementaria es una medida de otra y así no combinar dos medidas de calidad que aporten aproximadamente la misma información. Se obtiene el coeficiente de correlación entre todas las posibles combinaciones de cada una de las medidas que intervienen en el análisis.

Por último se realizan experimentos de utilidad con el objetivo de mostrar la efectividad de las medidas de calidad para predecir el rendimiento de un sistema. Para ello se emplean diferentes representaciones, graficas de dispersión (Scatter-plot), curvas DET entre otras.

En [35] se propone otra variante para calibrar un sistema de verificación de locutores. Se utiliza la duración de las señales como una medida de calidad para mejorar la calibración del 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 [36], pues se ha demostrado en [33] que este es menos sensible a las muestras de poca duración comparado 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 mejorar la calibración del sistema. Se emplean varias condiciones de duración, siempre entrenando con la totalidad de la señal y realizando la verificación con diferentes casos, el total del tiempo, 20 segundos o 5 segundos de la muestra. Por este motivo se entrenan diferentes parámetros de calibración para todas estas posibles condiciones de calidad modelando la relación entre el rango de valores de calidad y el proceso de calibración en una sola función. 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 de calidad.

# 6 Resultados experimentales

Las medidas de calidad explicadas en el epígrafe 4 no fueron diseñadas para aplicarlas al reconocimiento de locutores, sin embargo se han empleado en estos sistemas para medir la distorsión presente en la señal que se procesa. En la literatura aparecen resultados reportados en [5, 18] respecto a las tres primeras medidas descritas anteriormente, sin embargo de la medida HD no se reporta su aplicación en este tipo de sistemas.

En este epígrafe se describe la experimentación realizada a partir de la implementación de las mismas, con el objetivo de obtener una experiencia inicial en las formas de evaluar la relación que pueda existir entre dichas medidas de calidad con diversos tipos de ruido, la SNR, así como su relación con la eficacia del reconocimiento de locutores GMM-UBM.

La base de datos que se emplea es Ahumada v1.0 [37], que consta de 100 locutores de los cuales se utilizaron un pequeño grupo de 50 para el experimento: se tomaron 50 expresiones de una sesión microfónica para el entrenamiento y 50 de otra sesión microfónica para la prueba, las que fueron “ensuciadas” de manera electrónica con cuatro tipos de ruido: ruido blanco, ruido *street*, ruido *babble* y ruido música, con niveles de SNR de 0 dB, 5dB, 10 dB, 15 dB y 20 dB.

Además de esta prueba de reconocimiento con esta base de datos corrupta por ruido se realizó una prueba con los mismos locutores sin ruido para poder comparar luego los resultados obtenidos entre ambas pruebas.

Para realizar el análisis de los resultados se obtuvieron dos tipos de curvas: SNR vs Calidad y Score vs Calidad. Las medidas de calidad fueron mapeadas entre 0 y 1, como se explicó en el epígrafe 4.

## 6.1 Curvas de valor mapeado de calidad vs SNR para cuatro medidas de calidad y cuatro ruidos

A continuación se muestran en la figura 5, los resultados obtenidos para las cuatro medidas de calidad con respecto a los cuatro tipos de ruidos con sus cinco niveles de SNR. Para obtener estas curvas, se promedian los valores de calidad obtenidos para todas las muestras, luego de aplicarle la medida de calidad para un tipo de ruido y un nivel de SNR.

Con relación a las medidas de calidad que incluyen la kurtosis de los rasgos (KLPC y KCEP) se observa en general para todos los ruidos, un comportamiento no esperado pues la medida de calidad tiende a disminuir a medida que mejora la SNR, teniendo en cuenta que fueron utilizadas las funciones de mapeo propuestas.

De un análisis realizado por los autores, la posible causa de este comportamiento puede estar en la forma en que se mide y promedia la kurtosis de los rasgos LPC y CEP: por tramas [16]. Consideramos que los rasgos evaluados por tramas poseen muy poca gaussianidad, incluso en señales sin ruido, por lo

que proponemos que en estudios posteriores estos métodos sean revisados y se propongan nuevos métodos.

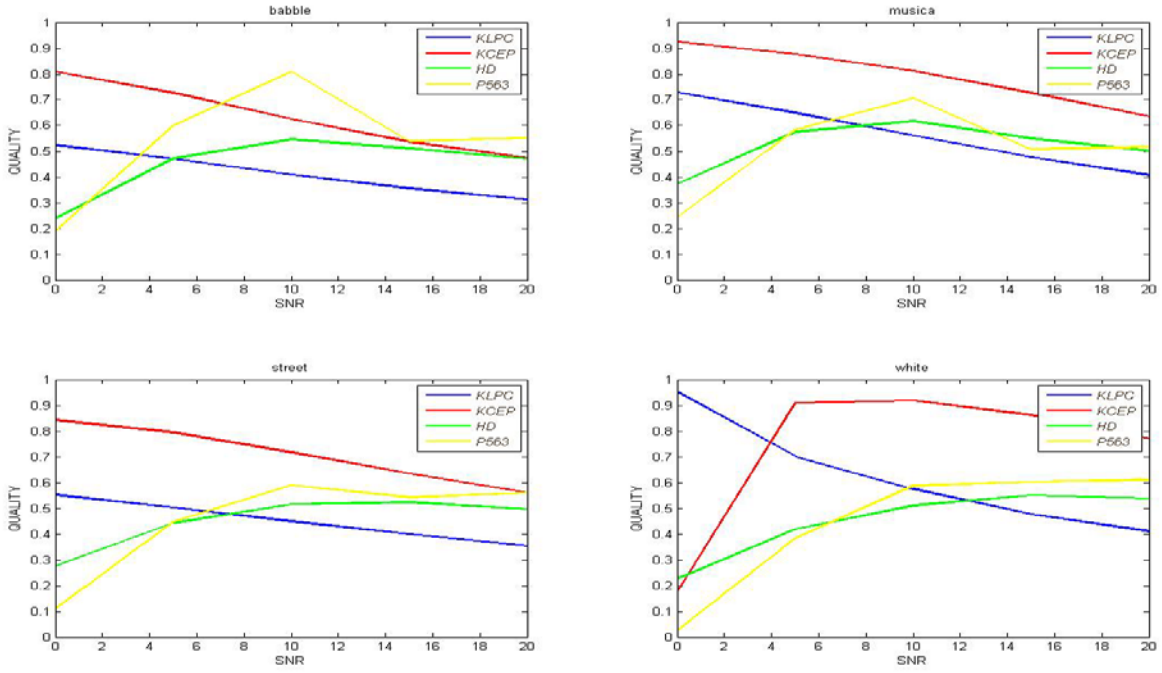


**Fig. 5.** Curvas de calidad vs SNR para ruido *babble*, música, *street*, blanco.

La medida HD no tiene un comportamiento ascendente con la SNR como debe suceder pues esta medida da idea de la armonicidad que existe en una señal de voz, es decir cuan marcados están los armónicos espectrales en ella. Para todos los ruidos, con baja SNR tiende a aumentar la HD pero al llegar a SNR= 10dB o mayor, el comportamiento cambia y la HD disminuye de manera contraria a cómo se espera, sobre todo ante el ruido babble y música. Este comportamiento se debe a que ambos son ruidos no estacionarios, sus variables estadísticas se modifican en el tiempo y no existirán componentes ruidosos en todas las bandas de frecuencia.

Usualmente la voz se asume como un proceso estacionario por tramas, pero al mezclarla con este tipo de ruidos el resultado será un proceso no estacionario. Los armónicos espectrales de una señal de voz están definidos como las zonas espectrales de la señal donde se concentran la mayor cantidad de potencia. Al mezclarse la señal con ruido no estacionario, estos se ven afectados de manera no uniforme para cada trama, al no existir uniformidad en cuanto a la potencia tanto en los valles como en los armónicos no se observará una relación clara entre la medida HD y la SNR para ruidos no estacionarios.

El ruido *street* es pseudo-estacionario por lo que afecta de manera algo más uniforme que los anteriores a la señal de voz. Por tal motivo la HD se corresponderá mejor con la SNR.

En el ruido blanco, totalmente estacionario, se afectan de forma más uniformes todas las componentes espectrales de la señal. En este caso los valles tienden a afectarse mucho más que los armónicos, y la potencia en los valles dará una idea de cuan distorsionada estará la señal. En este caso al igual que en el anterior, al aumentar la SNR la potencia en los armónicos disminuye pues existe menos energía de ruido en ellos, pero en los valles la potencia disminuye aún más porque allí solo existe la energía relacionada con el ruido. Es por este motivo que el comportamiento de la HD con relación a la

SNR en ruidos pseudo-estacionarios y estacionarios es mucho más preciso que en ruidos no estacionarios.

La P.563 tiene su peor comportamiento ante los ruidos *babble* y música, lo que se corresponde con las recomendaciones de su utilización por parte de la UIT-T, ya que esta medida no ha sido validada para estos ruidos y no debe utilizarse. En el caso de ruido *street* y blanco la relación directa entre esta medida y la SNR mejora considerablemente, correspondiéndose con las recomendaciones para su utilización.

### 6.2 Curvas de valores promedio de la medida de calidad vs valor medio de la puntuación de reconocimiento de locutores.

La figura 6 muestra la relación entre el valor obtenido de la medida de calidad y la puntuación de una prueba de verificación con 50 locutores. Esta representación pretende establecer un criterio de la capacidad discriminativa de un sistema de reconocimiento de locutores para distintas medidas de calidad. Si la medida de calidad es útil el sistema debería discriminar mejor si la calidad es más alta.

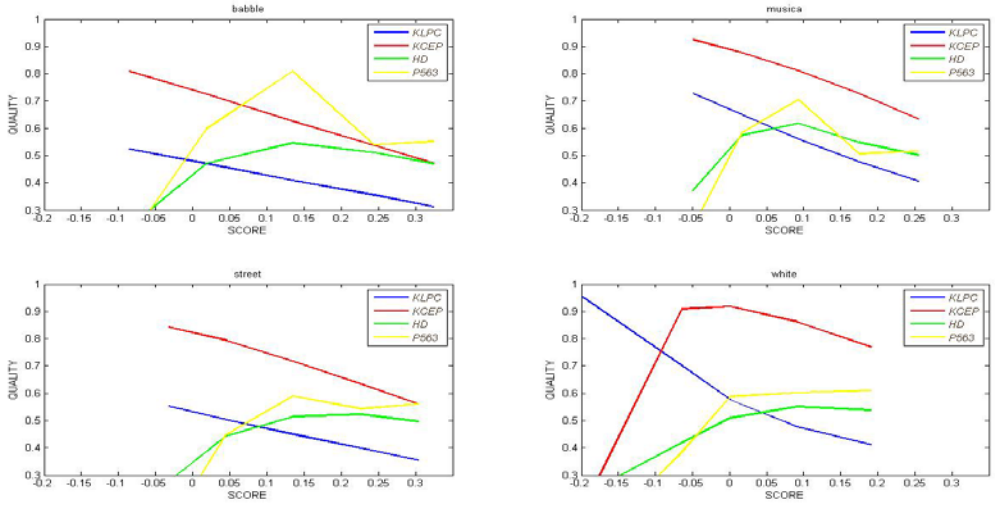


**Fig. 6.** Curvas de valores promedio de calidad vs puntuación de verificación promedio para los diferentes tipos de ruidos que se evalúan.

Se observa un comportamiento muy similar al de la figura 5, debido a la relación directa existente entre la SNR y la puntuación de la verificación.

A partir de ambos resultados, sugerimos diseñar experimentos con más locutores y revisar críticamente las medidas de kurtosis.

## 7 Conclusiones

Este reporte técnico pretende brindar una visión de las principales medidas de calidad utilizadas para la voz, así como de los intentos de aplicación de algunas de ellas en el reconocimiento de locutores, a partir de los resultados obtenidos en las experimentaciones que acompañan al reporte se pueden destacar ciertos aspectos de dichas medidas.

Es necesario señalar que las muestras utilizadas para realizar la experimentación fueron solo de 50 locutores y este factor influye evidentemente en los resultados. Nuestro objetivo fue probar los

métodos de evaluación y tratar de comprobar algunos criterios relacionados con la estacionariedad y correlación de los ruidos y su influencia en las medidas de calidad.

Se puede observar que existe una relación estrecha entre todas las medidas propuestas y la SNR. Es evidente que todas las medidas dependen en alguna forma de ella, al estar la SNR muy vinculada al resultado del reconocimiento de locutores, como un parámetro básico para definir la distorsión de una muestra y cuan confiable puede ser entonces la puntuación que se obtenga para ella.

Es de notar que las medidas relacionadas con la kurtosis (KLPC y KCEP) presentan un comportamiento no esperado.

La recomendación P.563 presenta un buen comportamiento en los casos en que ha sido validada por la UIT-T, por lo que sería útil indagar más en los 51 parámetros que este estándar evalúa para definir la calidad, a pesar de no estar diseñada para el reconocimiento de locutores.

La HD es una medida algo inestable y poco confiable para el reconocimiento de locutores pues varía mucho su desempeño de acuerdo con el tipo de ruido por lo que en un gran número de situaciones no se puede determinar su relación con la SNR y la puntuación.

Se describe además como las medidas de calidad han sido empleadas para diferentes fines en el reconocimiento de locutores, no solo para determinar su relación con la puntuación de una comparación. Es de notar como pueden aplicarse a diferentes momentos del reconocimiento, sobre todo en la determinación de la puntuación y su fusión, partiendo de la información que se utilice, ya sea de bajo o alto nivel o una combinación de ambas.

## 8 Recomendaciones y trabajo futuro

Es necesario realizar nuevas pruebas de reconocimiento de locutores con mayor cantidad de muestras para observar mejor la relación que debe existir entre la puntuación del reconocimiento y la medida de calidad, para determinar su validez o no.

Además como el comportamiento de las medidas está muy condicionado por el tipo y nivel de ruido que corrompe la señal de voz, es recomendable evaluar cómo cada tipo y nivel de ruido influye en los rasgos que se extraen de la señal y a su vez como esto influye en la medida de calidad. Esta evaluación puede hacerse de forma estadística, por cada rasgo.

A partir de estos resultados se propone:

1. Evaluar como debe ser el desempeño de las medidas de calidad bajo diferentes condiciones y tipos de ruido, a partir de cómo influye este en el comportamiento de los rasgos.
2. Proponer y evaluar nuevas medidas de calidad objetivas no intrusivas, diseñadas propiamente para el reconocimiento del locutor, ya sean diferentes formas de medir la SNR, el comportamiento de los formantes y el tono fundamental, el índice de modulación u otros.
3. Deben utilizarse y evaluarse dichas medidas de calidad en diversas etapas de un sistema de reconocimiento de locutores como los explicados en el epígrafe 5.

# Referencias bibliográficas


1. *IEEE Recommnded Pratice for Speech Quality Measurements*, in Audio and Electroacoustics, IEEE Transactions, 1969. **17**(3): p. 225-246.
2. Telecomunicaciones, U.I.d., *Methods for subjectivs determination of transmission quality P.800*, in *Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales.* 1996.
3. Telecomunicaciones, U.I.d., *Calidad de la Transmision telefonica. Prueba subjetiva de opinion. P.830*, in *Sector de normalizacion de las telecomunicaciones* 1998.
4. Benesty, J., M.M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*2007: Springer-Verlag New York, Inc.
5. Castro, A.H., *Fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz*, in *Dpto. de Ingeniería Informática* 2010, Universidad Autónoma de Madrid: Madrid.
6. Jonas Richiardi, A.D., *Evaluation of speech quality measures for the purpose of speaker verification*, 2008.
7. Loizou, P.C., *Speech Quality Asssessment*, in *Multimedia Analysis, Processing and Communications*, D.T. Weisi Lin, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, Haohong Wang, Editor 2011, springer.
8. Garcia-Romero, D., et al., *Using quality measures for multilevel speaker recognition.* Computer Speech & Language, 2006. **20**(2–3): p. 192-209.
9. Falk, T.H. and C. Wai-Yip, *Single-Ended Speech Quality Measurement Using Machine Learning Methods.* Audio, Speech, and Language Processing, IEEE Transactions on, 2006. **14**(6): p. 1935-1947.
10. Kondo, K., *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*2012: Springer.
11. Wang, S., A. Sekey, and A. Gersho, *An objective measure for predicting subjective quality of speech coders.* Selected Areas in Communications, IEEE Journal on, 1992. **10**(5): p. 819-829.
12. Telecomuniaciones, U.I.d., *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.*, in *Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales. Métodos de evaluación objetiva y subjetiva de la calidad.*2001.
13. Kitawaki, N., H. Nagabuchi, and K. Itoh, *Objective quality evaluation for low-bit-rate speech coding systems.* Selected Areas in Communications, IEEE Journal on, 1988. **6**(2): p. 242-248.
14. Itakura, F. and T. Umezaki. *Distance measure for speech recognition based on the smoothed group delay spectrum*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87*. 1987.
15. Vaseghi, S.V., *Advanced Digital Signal Processing and Noise Reduction*. 4th ed2008: John Wiley&Sons.
16. Hirsch, H.G. and C. Ehrlicher. *Noise estimation techniques for robust speech recognition*. in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. 1995.
17. Richiardi, J. and A. Drygajlo. *Evaluation of speech quality measures for the purpose of speaker verification*. in *Proc. Odyssey 2008: The Speaker and Language Recognition Workshop*. 2008.
18. Telecomunicaciones, U.I.d., *Método basado en un solo extremo para la evaluación objetiva de la calidad vocal en aplicaciones de telefonía de banda estrecha P.563*, in *Serie P: Calidad de transmision telefónica, instalaciones telefónicas y redes locales. Aparatos para mediciones objetivas.*2004.
19. Harriero, A., et al., *Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification*, in *Proceedings of the Third International Conference on Advances in Biometrics*2009, Springer-Verlag: Alghero, Italy. p. 434-442.
20. Falk, T.H. and C. Wai-Yip. *Enhanced Non-Intrusive Speech Quality Measurement Using Degradation Models*. in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. 2006.
21. Gray, P., M.P. Hollier, and R.E. Massara, *Non-intrusive speech-quality assessment using vocal-tract models.* Vision, Image and Signal Processing, IEE Proceedings -, 2000. **147**(6): p. 493-501.
22. DeCarlo, L.T., *On the meaning and use of kurtosis.* Psychological Methods, 1997. **2**(3): p. 292-307.
23. Muro, A.G., *utilizacion de medidas de calidad de la señal de voz para compensacion de variabilidad inter-sesion en reconocimiento de locutor*, in *Dpto. de Ingeniería Informática, ATVS -Grupo de Reconocimiento Biométrico* 2012, Universidad Autónoma de Madrid.
24. 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.* Speech Commun., 1993. **12**(3): p. 247-251.

25. Yu, A.-T. and H.-C. Wang, *New speech harmonic structure measure and its applications to speech processing.* The Journal of the Acoustical Society of America, 2006. **120**(5): p. 2938-2949.
26. Kelly, F., A. Drygajlo, and N. Harte, *Compensating for Ageing and Quality variation in Speaker Verification*, in *INTERSPEECH*2012, ISCA.
27. Garcia-Romero, D., et al. *On the Use of Quality Measures for Text-Independent Speaker Recognition*. in *SPEAKER AND LANGUAGE RECOGNITION WORKSHOP (ODYSSEY)*. 2004.
28. Bigun, J., et al. *Multimodal biometric authentication using quality signals in mobile communications*. in *Image Analysis and Processing, 2003.Proceedings. 12th International Conference on*. 2003.
29. Fierrez-Aguilar, J., et al., *Discriminative multimodal biometric authentication based on quality measures.* Pattern Recognition, 2005. **38**(5): p. 777-779.
30. Malayath, N., et al., *Data-Driven Temporal Filters and Alternatives to GMM in Speaker Verification.* Digital Signal Processing, 2000. **10**(1–3): p. 55-74.
31. Richiardi, J., A. Drygajlo, and P. Prodanov, *Confidence and reliability measures in speaker verification.* Journal of the Franklin Institute, 2006. **343**(6): p. 574-595.
32. Villalba, J., et al., *Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures*, in *Advances in Speech and Language Technologies for Iberian Languages*, D. Torre Toledano, et al., Editors. 2012, Springer Berlin Heidelberg. p. 1-10.
33. Dehak, N., et al., *Front-End Factor Analysis for Speaker Verification.* Trans. Audio, Speech and Lang. Proc., 2011. **19**(4): p. 788-798.
34. Perez-Gomez, S., et al. *Modelos de regresión logística a nivel de puntuaciones para incorporar calidad en la verificación de locutor*. in *Actas de las V Jornadas de Reconocimiento Biométrico de Personas*. 2010. Huesca, España.
35. Mandasari, M., et al., *Quality Measure Functions for Calibration of Speaker Recognition System in Various Duration Conditions.* Audio, Speech, and Language Processing, IEEE Transactions on, 2013. **PP**(99): p. 1-1.
36. Dehak, N., et al. *Cosine Similarity Scoring without Score Normalization Techniques*. in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop*2010.
37. Ortega-Garcia, J., J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, *AHUMADA: A large speech corpus in Spanish for speaker characterization and identification.* Speech Communication, 2000. **31**(2–3): p. 255-264.

RT_058, enero 2014
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*