UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz

-PROYECTO FIN DE CARRERA-

Alberto Harriero Castro
Febrero de 2010

Fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz

AUTOR: Alberto Harriero Castro
TUTOR: Daniel Ramos Castro

ATVS – Grupo de Reconocimiento Biométrico
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero de 2010

Palabras clave

Sistemas de reconocimiento de locutor, sistemas forenses, calidad, indicador de degradación, utilidad, P.563, UBML , kurtosis, skewness.

Resumen

En este trabajo se estudian distintos métodos para medir la calidad de la señal de voz, y se evalúa su utilidad de cara a su aplicación a sistemas de reconocimiento automático de locutor. Algunos de estos métodos han sido ya probados en otros estudios sobre la materia, otros están basados en técnicas de monitorización de la calidad del servicio en la red telefónica, y el resto se proponen en este trabajo. De todos ellos se realiza un análisis exhaustivo, experimentándose sobre dos bases de datos proporcionadas por el NIST (NIST SRE 2006 y SRE 2008), comúnmente utilizadas en diversos estudios del estado del arte, más una base de datos forense (Ahumada III) con conversaciones reales que presentan una gran variabilidad de la calidad de la señal de voz. Todos los experimentos se llevan a cabo con 3 sistemas diferentes implementados por el ATVS – Grupo de Reconocimiento Biométrico, que utilizan las últimas técnicas en reconocimiento automático de locutor con compensación de variabilidad intersesión.

Al inicio del trabajo se realiza una introducción al reconocimiento biométrico y los sistemas de reconocimiento de locutor. Más adelante se presentan los conceptos básicos sobre calidad en reconocimiento biométrico, los métodos empleados para medir la calidad de la voz al igual que se presentarán las medidas de calidad que se estudiarán en el trabajo. A continuación se explicará la metodología experimental empleada, la cual se basa en estudios previos en la materia, y sobre esta se propone un nuevo método de análisis que complementa los métodos utilizados hasta el momento: el estudio de los indicadores de degradación, que consiste en la obtención de gráficas (que llamaremos gráficas de “Rendimiento vs Magnitud”) que representan la relación entre la magnitud de los parámetros que se quieren estudiar como medidas de calidad (indicadores de degradación) y el rendimiento de los sistemas. En la sección de experimentos se presentan los resultados obtenidos rigiéndose por la mencionada metodología.

Por último se presentan las conclusiones del proyecto y las líneas de trabajo futuro.

Abstract

In this work we study several methods to measure the voice signal quality as well as their utility as an application to speaker recognition systems is evaluated. Some of these methods have been already tested in other studies on this field, some others have been tested to monitoring quality of service on telephony networks and the rest of them are proposed on this work. An deep analysis is performed for all of these methods; two databases provided by the NIST are used (SRE 2006 and SRE 2008), commonly used on the state of the art, plus a forensic database (Ahumada III) which contains real conversations with a high level of quality variability of the voice signal. All experiments are carried out with three different systems developed in the ATVS – Biometric Recognition Group, which include last techniques on automatic speaker recognition with intersession variability compensation.

At the beginning of this work an introduction to biometrics and speaker recognition systems is made. Next, basic concepts on quality in biometrics are introduced, as well as the common methods of quality estimation, and the quality measures studied. Then the methodology used to develop the experiments is presented, which is based on previous work on this field, and it is completed with a novel method to test the utility of a quality measures that adds complementary information to all methods studied so far: the degradation indicators study, which consists of obtaining graphics that represent the relation between the parameters which are intended to be studied as quality measures (degradation indicators) and the performance of the systems. Results are presented on the experiments section.

Finally the project conclusions are drawn and future lines of work are presented.

Agradecimientos

En primer lugar quiero agradecer a mi tutor Daniel Ramos la oportunidad de desarrollar mi Proyecto de Fin de Carrera en un grupo puntero en el reconocimiento biométrico como es el ATVS. También quiero agradecerle todo el apoyo, la dedicación y el interés que ha prestado durante todo el desarrollo de este proyecto, sin los cuales estoy convencido de que la calidad no sería la misma.

También me gustaría agradecer a todo el equipo del grupo ATVS, y especialmente a aquellos con los que he compartido el día a día, toda la ayuda que me han prestado, sin la cual el desarrollo de este proyecto hubiera sido aún más duro de sacar adelante: Alejandro Abejón, Javier González, Javier Franco, Ignacio López, Fernando García, Verónica Peña, Ismael Mateos, Sergio Pérez y Sergio Lucas. Gracias a todos.

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica superior de la Universidad Autónoma de Madrid.

Índice

Índice de figuras

Índice de tablas

1. INTRODUCCIÓN

1.1 Motivación del PFC

Las tecnologías de reconocimiento biométrico han experimentado una notable mejora en los últimos años que ha propiciado una gran expansión y aceptación de los mismos en distintos tipos de aplicaciones (seguridad, domótica, comercio electrónico, sistemas forenses, etc.).

Dentro de los rasgos biométricos, la voz siempre ha sido estudiada como un rasgo bien diferenciado de la mayoría. Entre otras peculiaridades, la voz destaca por estar sometida a muchas fuentes de variabilidad: al contrario que otros rasgos biométricos, se ve influido tanto por las características físicas del individuo como por su comportamiento. Además, frecuentemente se trabaja con señales de voz transmitidas a través de una red telefónica, lo cual implica diversos factores que pueden degradar la señal. Por último otros factores como el entorno o el dispositivo de registro también pueden influir en la fidelidad de una muestra. Estos últimos tienen especial importancia en casos forenses, ya que en estos nunca se controla el entorno ni el dispositivo de registro [Ramos et. al, 2008].

Tanto en voz como en el resto de rasgos biométricos se han desarrollado métodos cuyo objetivo es compensar esta variabilidad de modo que de las muestras biométricas se obtenga información lo más fiel posible a la fuente. Dentro de estos métodos se distinguen dos grandes familias: las guiadas con información previa (que tratan de modelar las variabilidades con grandes cantidades de datos) y las basadas en información sobre los rasgos biométricos , que consiste básicamente en el estudio de la calidad de las muestras biométricas. De la primera de las familias existen diferentes métodos que llevan aplicándose desde en la última década con gran éxito en todos los tipos de sistemas biométricos (incluidos los de voz) [Deller *et al.* 1999]. La calidad sin embargo ha experimentado un avance mucho mayor en los sistemas basados en imagen (la mayoría) que en los sistemas de reconocimiento de locutor.

El organismo estadounidense NIST realiza un *workshop* bianual sobre calidad [NIST QUALITY WORKSHOP] en el cual se comparten los últimos avances en la materia. Las aportaciones en materia de calidad en voz han sido mínimas o nulas en este tipo de foros científicos, lo cual parece una paradoja si tenemos en cuenta todos los factores de variabilidad que influyen sobre la señal de voz, y que ésta a su vez constituye el rasgo biométrico más diferenciado del resto.

Los métodos de compensación de calidad se componen de dos principales pasos que son igualmente importantes. El primero de ellos sería la medición de la calidad, que será tanto mejor cuanto mayor sea la precisión con la que se mide. Y el segundo paso consiste en utilizar las mediciones para compensar la calidad de la muestra biométrica. Es importante resaltar, que la efectividad del segundo se basa totalmente en el primero: para poder gestionar la calidad, primero hay que ser capaz de medirla correctamente. El criterio utilizado para determinar si una medida es capaz de determinar la calidad de una muestra ha sido la utilidad, que se define como la capacidad de una medida para predecir el rendimiento de un sistema para un valor dado de calidad. Este criterio se ha tomado de otros estudios en la materia [M1/05-0306; Alonso *et al.* 2008; Grother *et al.* 2007].

El objetivo de este PFC es realizar un análisis en profundidad de diferentes métodos de estimación de calidad de la voz, para conocer la precisión con la cual estos determinan la calidad de las muestras de voz. Algunos de los métodos ya habían sido estudiados en el estado del arte [Gª Romero *et.al.*, 2005; Richiardi y Drygajlo, 2007], y aquí se amplía el análisis sobre ellos. El resto de métodos proceden de aplicaciones a Quality of service (QoS) [ITU-T P.563] que nunca se han aplicado en este ámbito, o bien se proponen en este trabajo, constituyendo una contribución del mismo.

Este análisis pretende ser una sólida base para una línea de investigación que ya está en marcha en el ATVS – Grupo de Reconocimiento Biométrico, y cuyo objetivo será seguir haciendo avances y conseguir posicionarse como un referente en el análisis de calidad en sistemas de reconocimiento biométrico. Los primeros avances, fruto de este trabajo se presentaron en el ICB (*International Conference on Biometrics)* en junio de 2009 (ver Anexo B).

1.2 Objetivo y enfoque

El objetivo de este trabajo es realizar un análisis exhaustivo de varios métodos de estimación de calidad, obteniendo la mayor cantidad de información posible, de modo que sirva como base para futuros experimentos de compensación de calidad. Este objetivo lo desglosamos a su vez en los puntos que siguen:

1. Estudiar el estado del arte. Recopilar las últimas tecnologías utilizadas en SRL, los métodos utilizados para analizar medidas de calidad y los métodos utilizados en voz para estimar la calidad de la señal.
2. Implementación de medidas de calidad. Tomando como base estudios de calidad en voz (tanto de SRL como de QoS), implementar sus métodos de estimación de calidad de la voz para realizar un análisis exhaustivo de los mismos.
3. Proponer nuevas medidas de calidad y así realizar una aportación en la materia que permita mejorar en un futuro los resultados de la compensación utilizando la calidad de la voz.
4. Evaluar las medidas de calidad basándonos en protocolos recomendados por NIST [M1/05-0306] que han sido utilizados en otros estudios de calidad en biometría.

Debido a la escasez de estudios previos en la materia, el trabajo de documentación realizado en este trabajo juega un papel fundamental, ya que servirá como base para seleccionar los métodos de estimación de calidad a estudiar. Dicha documentación se basará en gran medida en los estudios existentes sobre estimación objetiva de la calidad subjetiva (principalmente [ITU-T P.563]), cuyo fin es determinar la calidad media que otorgaría un usuario a una locución en una red de telefonía.

Ajustándose a los objetivos planteados, se han realizado una serie de experimentos sobre una base experimental sólida, cuyas características principales se citan a continuación:

•**Distintas bases de datos**: se utilizarán dos bases de datos proporcionadas por el NIST (SRE 2006 y SRE 2008) [NIST SRE 2006; NIST SRE 2008] que cuentan con numerosos estudios en SRL. Ambas contienen locuciones de origen telefónico, lo que permitirá realizar una comparación entre bases de datos en habla telefónica. Además, SRE 2008 contiene locuciones de origen microfónico (sin transmitirse por una red telefónica) que nos permitirá observar la

influencia de la calidad con este tipo de habla, al igual que en mezclas con distintos tipos de habla (p.ej. modelos de habla telefónica y locuciones de testeo microfónicas). Además, se experimentará con una base de datos forense registrada por el departamento de Imagen y Acústica de la Guardia Civil llamada Ahumada III [Ramos *et. al.,* 2008], en la cual las condiciones de variabilidad de la calidad son extremas.

•**Distintos sistemas**: se probará como influye la calidad de las muestras biométricas en tres sistemas diferentes, los tres implementados por el ATVS – Grupo de Reconocimiento Biométrico que incorporan las últimas técnicas del estado del arte, a saber: Modelos de Mezclas Gausianas o *Gaussian Mixture Models* (GMM), Máquinas de Vectores Soporte o *Support Vector Machines* con un kernel GLDS (SVM-GLDS) y por último el sistema híbrido de Máquinas de Vectores Soporte basada en Súper Vectores (SVM-SV). Esto, además de permitir contrastar los resultados entre distintos sistemas, da una idea de cuánto puede variar la influencia de la calidad cuando se trata de distintos sistemas.

•**Distintos métodos de estimación de calidad**: se estudiarán métodos ya utilizados en otros estudios en la materia, lo que permitirá contrastar la efectividad de otros métodos que se proponen en este trabajo.

Basándonos en trabajos previos en la materia, los experimentos aportan información sobre los distintos métodos atendiendo a dos propiedades principales:

•**Experimentos de utilidad**: es el método más utilizado para determinar la efectividad de un método de estimación o medida de calidad. Consiste en saber si una medida es capaz de predecir de manera aproximada el impacto que tendría la calidad en el rendimiento de un sistema de reconocimiento biométrico.

•**Experimentos de correlación**: consiste en conocer el grado en el que se complementan las informaciones aportadas por distintos métodos, de modo que puedan combinarse en un futuro de manera óptima.

A estos dos tipos de experimentos habría que añadir un tercer tipo que se propone en este trabajo: el **estudio de los indicadores de degradación**. Su objetivo es aproximar la relación existente entre la magnitud de los parámetros que indican la degradación de la voz (indicadores de degradación) y el rendimiento de los sistemas.

1.3 Contribuciones originales del PFC

A continuación se citan las principales contribuciones que aporta este trabajo:

•Nuevas medidas de calidad

En este punto habría que destacar dos grupos:

- Los nuevos métodos propuestos de estimación de calidad de la voz, que son dos: la Similitud con el Modelo de Habla Universal o *Universal Background Model Likelihood* (UBML) y la estimación de la Relación Señal a ruido por filtrado de Wiener. Ambas serán explicadas en la sección 3. Como podrá comprobarse, la primera de ellas

presenta unas cualidades excepcionales, destacando entre todas las medidas estudiadas.

- Métodos de estimación de calidad en QoS que no habían sido probados en sistemas de reconocimiento de locutor. Serán vistas en la sección 3.

•Nuevos métodos de análisis de medidas de calidad

Estudio de los indicadores de degradación. Este método de análisis propuesto en este trabajo aportará información adicional sobre las medidas de calidad (sección 3).

•Estudios experimentales

- Estudio de la influencia de la calidad en sistemas de reconocimiento de locutor con bases de datos telefónicas.
- Estudio de la influencia de la calidad en sistemas de reconocimiento de locutor con bases de datos microfónicas.
- Estudio de la influencia de la calidad en sistemas de reconocimiento de locutor en cruces de micrófono y teléfono.
- Estudio de la influencia de la calidad en sistemas de reconocimiento de locutor con bases de datos forenses.

•Datos estadísticos sobre las medidas de calidad estudiadas

- Histogramas de los valores de las medidas de calidad estudiadas (Anexo A).
- Valores de correlación entre las distintas medidas, que permitirá combinarlas de manera óptima en futuros estudios de compensación de calidad.

2. FUNDAMENTOS DE SISTEMAS DE RECONOCIMIENTO BIOMÉTRICO

En este capítulo se presentarán los aspectos más importantes para comprender el funcionamiento de un sistema de reconocimiento biométrico, para después entrar más en detalle con el funcionamiento de los sistemas de reconocimiento de locutor, repasando al mismo tiempo los trabajos más relevantes del estado del arte.

2.1 Rasgos biométricos

La finalidad del reconocimiento biométrico es la clasificación de un sujeto en base a uno o más rasgos del mismo (rasgos biométricos). Estos rasgos pueden venir determinados bien por sus propiedades físicas, o por su manera de actuar.

Para que un rasgo biométrico sea considerado como tal, debe cumplir los siguientes requisitos [Maltoni *et al.*, 2003]:

•**Universalidad**: debe ser un rasgo común a toda la población de interés.

•**Distintividad**: dos personas cualesquiera deben ser lo suficientemente diferentes para poder ser diferenciadas.

•**Estabilidad**: debe ser un rasgo cuya variación a lo largo del tiempo sea lo suficientemente pequeña como para poder ser identificado.

•**Evaluabilidad**: debe poder ser caracterizado de forma cuantitativa.

Además, en aplicaciones reales se considera que un sistema basado en un rasgo biométrico determinado debe cumplir las siguientes características:

•**Rendimiento**: hace referencia a la precisión y velocidad con las que opera el sistema.

•**Aceptabilidad**: los usuarios del sistema deben estar dispuestos a utilizar dichos rasgos para ser reconocidos.

•**Fraude**: el sistema debe ser robusto ante posibles ataques. Esta característica es especialmente importante en sistemas forenses, ya que su finalidad es proporcionar una evidencia fiable sobre la identificación de un individuo.

Los rasgos biométricos que cumplen en mayor o menor medida los requisitos mencionados y que por tanto están considerados como rasgos biométricos son: ADN, dinámica de tecleo, escáner de retina, firma, forma de caminar, geometría de la mano, huella dactilar, iris, olor, oreja, rostro, termograma facial, venas de la mano y voz.

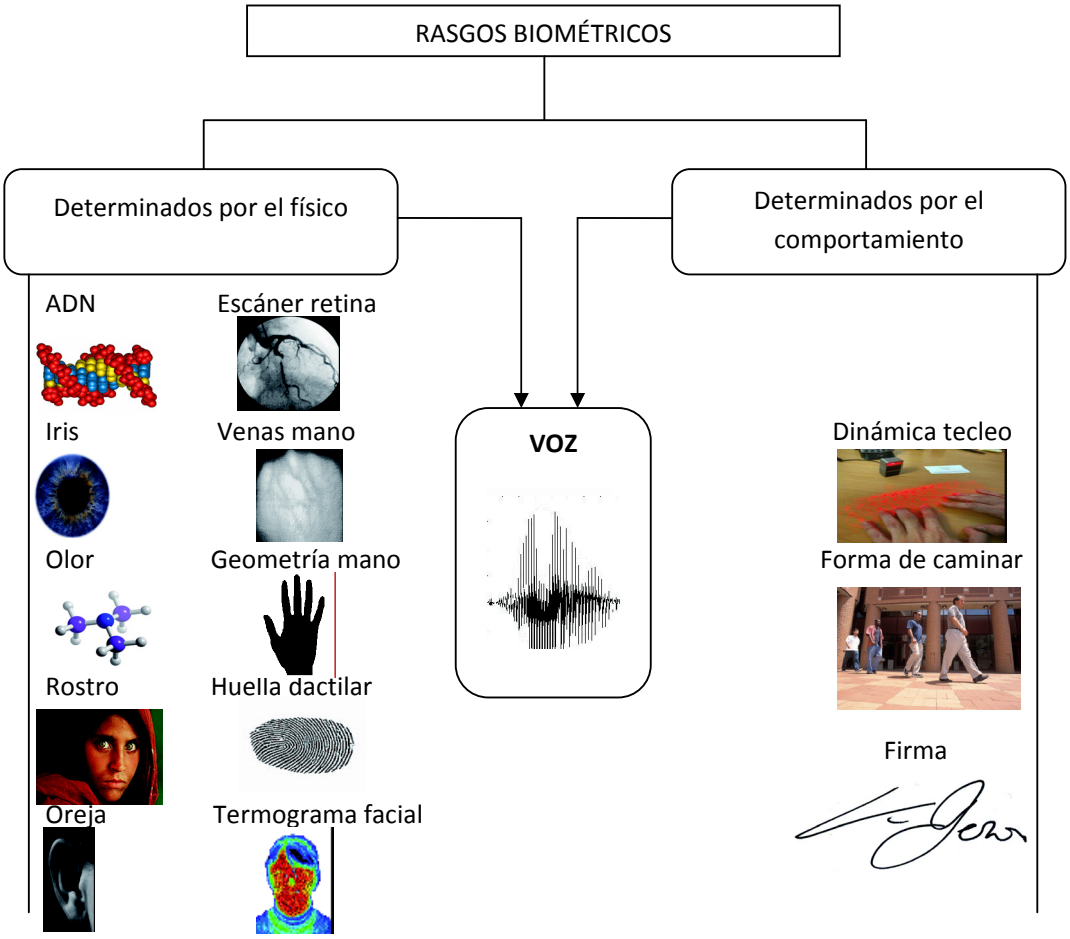


Fig 2.1: *rasgos biométricos contemplados en el estado del arte*.

	Universalid	Distintivida	Estabilidad	Evaluablilid	Rendimient	Aceptabilid	Fraude
ADN	A	A	A	B	A	B	B
Dinámica del tecleo	B	B	B	M	B	M	M
Escáner de retina	A	A	M	B	A	B	B
Firma	B	B	B	A	B	A	A
Forma de caminar	M	B	B	A	B	A	M
Geometría de la mano	M	M	M	A	M	M	M
Huella dactilar	M	A	A	M	A	M	M
Iris	A	A	A	M	A	B	B
Olor	A	A	A	B	B	M	B
Oreja	M	M	A	M	M	A	M
Rostro	A	B	M	A	B	A	A
Termograma facial	A	A	B	A	M	A	B
Venas de la mano	M	M	M	M	M	M	B
Voz	M	B	B	M	B	A	A

Tabla 2.1: *resumen de las características de todos los rasgos biométricos*.

Las características más destacables de la voz son las que a continuación se citan:

•Determinada por el físico y por el comportamiento: aunque la más determinante sea la primera, la voz de cualquier individuo puede sufrir cambios notables ya sea por factores externos (enfermedades, emociones, entorno) o voluntariamente.

•Rasgo muy aceptado y accesible: debido al uso cotidiano de la voz, es fácil tener acceso a este rasgo físicamente o a distancia, y su registro no suele ser interpretado como una amenaza para la privacidad de la gente, por lo que es comúnmente aceptado.

•Baja distintividad y estabilidad. Estas dos características, son en parte consecuencia de ser un rasgo dependiente del comportamiento. Esto hace más difícil el reconocimiento de voz con respecto a otros rasgos biométricos.

2.2 Fundamentos de los sistemas de reconocimiento biométrico

2.2.1 Funcionamiento

En una primera aproximación podemos describir un sistema de reconocimiento biométrico como un clasificador de patrones que toma como entrada una muestra de un rasgo biométrico de un individuo, y le asigna una determinada identidad. En la siguiente figura se recoge un esquema a alto nivel del funcionamiento de un sistema de reconocimiento biométrico. En él se pueden observar los diferentes bloques funcionales que lo compondrían.

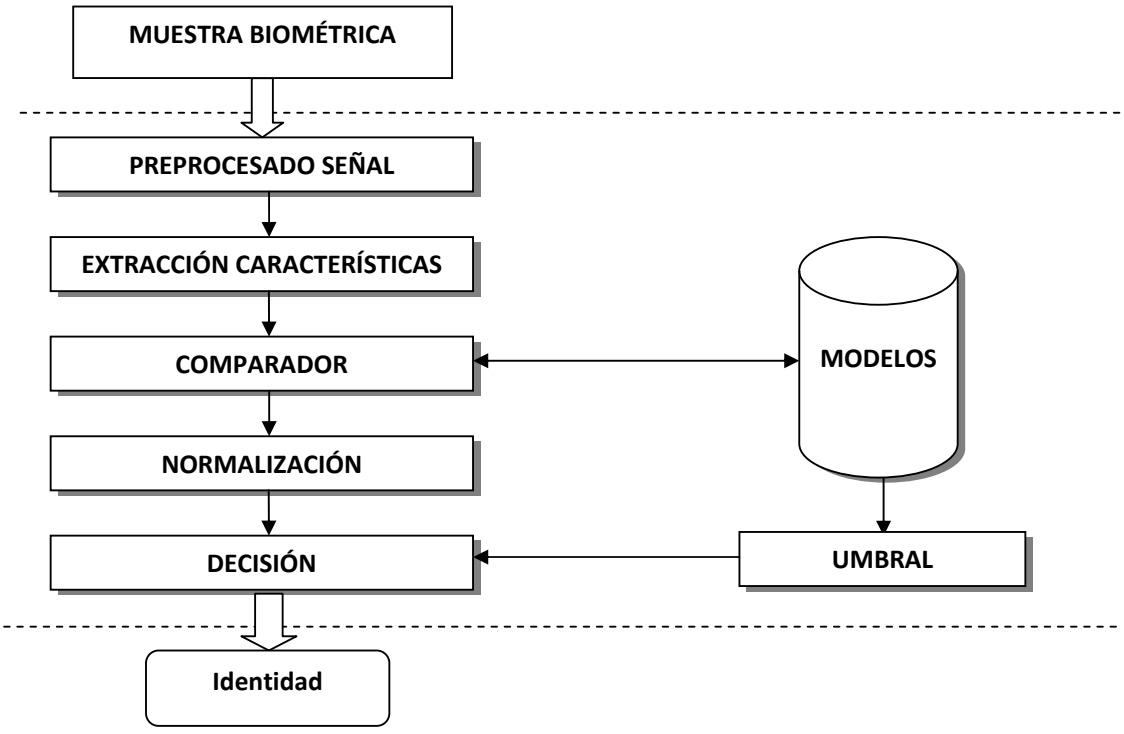


Fig 2.2: *esquema a alto nivel del funcionamiento de un sistema de reconocimiento biométrico.*

En primer lugar se sometería la muestra a un procesado de la señal que permitirá extraer la mayor información de la misma en la siguiente fase. A continuación se extraen una serie de

parámetros de la muestra biométrica, que contendrán la información sobre la identidad del individuo. Dichos parámetros son comparados con uno o varios modelos (dependiendo del número de posibles identidades), produciendo una puntuación de similitud entre cada uno de los modelos y la muestra, que tendrá que ser normalizada. Con esta última puntuación, y teniendo en cuenta un umbral de decisión, en el último bloque se decide la identidad asignada a la muestra biométrica.

2.2.2 Modos de operación de un sistema de reconocimiento biométrico

Una primera clasificación diferenciaría dos tipos de reconocimiento: reconocimiento positivo y negativo. El positivo persigue demostrar que una persona es quien afirma ser, mientras que el negativo consiste en demostrar que a una persona no le corresponde una determinada identidad.

Dependiendo del contexto de la aplicación, un sistema de reconocimiento biométrico puede funcionar en dos modos:

•**Verificación**: su finalidad es demostrar que la identidad que afirma poseer un individuo corresponde al mismo. Se llevan a cabo 2 comparaciones: una con el modelo de la supuesta identidad, y otra con un modelo que representa a una población de interés. De esta manera, la verificación equivale a una clasificación binaria entre un individuo o "el resto de la población". Suele utilizarse en reconocimiento positivo para evitar que dos personas utilicen la misma identidad.

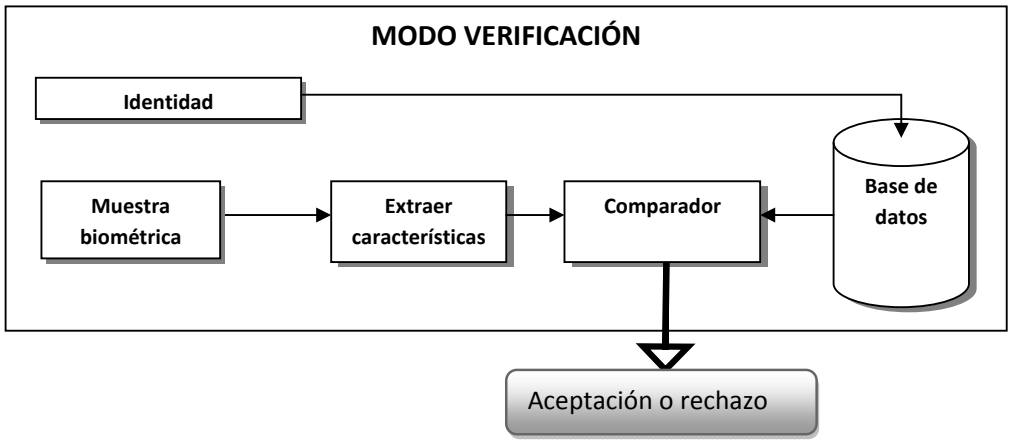


Fig 2.3*: esquema de un sistema de reconocimiento biométrico en modo de verificación.*

•**Identificación**: el sistema asigna al individuo una de las identidades almacenadas en su base de datos. Dentro de este pueden distinguirse dos tipos:

- Conjunto cerrado: el individuo siempre pertenece a alguna de las clases almacenadas en el sistema.
- Conjunto abierto: el individuo puede no pertenecer a ninguna de las clases registradas en el sistema.

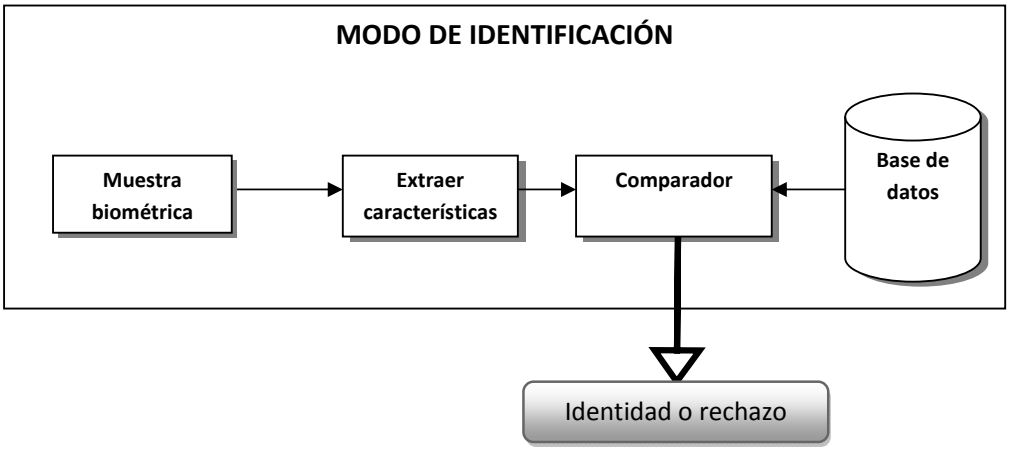


Fig 2.4*: esquema de un sistema de reconocimiento biométrico en modo de identificación.*

Cabe destacar un tercer modo de operación: el **modo de registro**. Constituye una fase previa a los dos citados anteriormente, y consiste en dar de alta a un nuevo usuario del sistema. Para ello es necesario obtener una o varias muestras biométricas del mismo para generar un modelo del mismo.

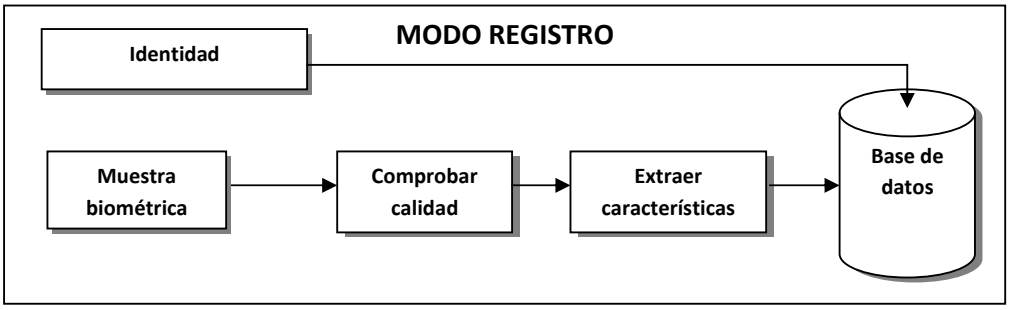


Fig 2.5*: esquema de un sistema de reconocimiento biométrico en modo de registro.*

2.2.3 Tipos de aplicaciones

Existen tres grandes campos en los que se pueden aplicar los sistemas de reconocimiento biométrico [Gonzalez-Rodriguez *et al*. 2007]:

•**Aplicaciones comerciales**: control de acceso, protección de datos electrónicos, cajeros automáticos, dispositivos electrónicos (ordenadores, PDAs, teléfonos móviles).

•**Aplicaciones gubernamentales:** documentos de identidad, seguridad social, control de fronteras.

•**Aplicaciones forenses**: puede aplicarse para identificación de terroristas, determinación de parentescos, identificación de cadáveres, evaluación de evidencias, etc.

Hasta el día de hoy, en las aplicaciones comerciales se utilizaban otros métodos de autenticación como claves o tarjetas magnéticas, mientras que las aplicaciones forenses eran confiadas a expertos que determinaban el peso de evidencias biométricas (ADN, huellas, voz) conforme a ciertos criterios.

Las grandes ventajas de los sistemas biométricos han permitido abrirse paso en aplicaciones comerciales y gubernamentales, teniendo un peso importante a día de hoy. En el ámbito forense actualmente se le otorga fiabilidad limitada en comparación con el análisis por parte de expertos humanos.

2.3 Limitaciones de los sistemas biométricos

Existen diversos factores que limitan el rendimiento de los sistemas de reconocimiento biométrico. Aquí citamos algunos de ellos:

•**Actitud del individuo**. Una actitud no cooperativa puede dar como resultado muestras que no reflejen las características del individuo. Un claro ejemplo es el fraude.

•**Procesos de extracción**. Existen ciertos factores que pueden influir en la fidelidad de la muestra biométrica: el entorno de la extracción, precisión del dispositivo de captura o el control sobre el individuo.

•**Envejecimiento**: con el paso del tiempo los rasgos biométricos se modifican. De todos ellos la voz es uno de los rasgos más propenso al cambio. Otros como cicatrices u operaciones estéticas (en el caso de imágenes). En el caso de la voz, problemas de salud (resfriados o voz afónica), o estados emocionales alterados, pueden producir modificaciones en la voz de un sujeto.

Todos estos factores tienen en común que dan como resultado una reducción de la información del rasgo biométrico en la muestra. Algunos de estos factores son controlables y se intentan controlar de distintas maneras. Otros en cambio son más difíciles de controlar. En caso de no poder controlarlos conviene al menos poder medirlos, con el fin de gestionarlos para compensarlos en la medida de lo posible.

El estudio de la calidad en biometría tiene como objetivo la obtención de parámetros que determinen ciertos factores reductores de la información de las muestras biométricas.

2.4 Sistemas de reconocimiento de locutor. Estado del arte.

Pese a sus limitaciones frente a otros rasgos biométricos, la voz es uno de los que tiene mayor peso entre todos los utilizados actualmente. Esto es por dos principales razones [Reynolds *et al.* 2004]:

•Es **comúnmente aceptado** por la sociedad. Al tratarse de un acto tan cotidiano, no se considera una amenaza para la privacidad de las personas.

•Gracias a **la red telefónica** se puede hacer uso del mismo desde prácticamente cualquier punto del planeta.

En este apartado se define con más detalle las características de la voz como rasgo biométrico, sus fundamentos, sus aplicaciones y por último los factores que la limitan, para acabar introduciendo la calidad como herramienta de compensación de dichos factores.

2.4.1 Información de la identidad en la señal de voz

La producción de la voz es un proceso bastante complejo que viene determinado por las características físicas del individuo (su tracto vocal) y otros factores aprendidos como el nivel de educación, acento, o el contexto sociolingüístico. En estos mismos factores se basa el sistema de percepción humano para reconocer personas por su voz. Los sistemas automáticos han tomado esta idea extrayendo información de la señal de voz a distintos niveles [Reynolds 2003]:

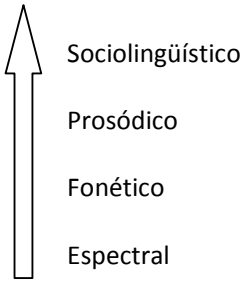


Fig 2.6: *niveles de información en la señal de voz.*

•**Nivel acústico o espectral**. La información se obtiene del espectro de la señal, la cual está directamente relacionada con la configuración dinámica del tracto vocal. Para ello se toman ventanas de duración muy corta (decenas de milisegundos) en las cuales se considera que la configuración del tracto vocal no varía, y se extraen ciertos parámetros de cada una de ellas.

•**Nivel fonético**. Cada persona hace un uso de los fonemas y silabas diferente. Aprovechando esta propiedad, es posible extraer parámetros que caractericen el habla de una persona basándose en la pronunciación de dichos fonemas.

•**Nivel prosódico**. La prosodia analiza y representa aquellos elementos de la expresión oral como los tonos y la entonación. Su manifestación concreta en la producción de la palabra se asocia a las variaciones de la frecuencia fundamental, de la duración y de la intensidad que constituyen los parámetros prosódicos.

•**Nivel sociolingüístico**. Factores como el nivel de educación, el contexto sociolingüístico o el origen de un sujeto determinan el uso del lenguaje de las personas.

2.4.1 Aplicaciones

En primer lugar debemos distinguir dos tipos de aplicaciones: dependientes de texto e independientes de texto.

•**Dependientes de texto**. En este tipo de sistemas el usuario debe pronunciar un conjunto de palabras que el sistema conoce. Es habitual que las palabras consistan en un número PIN de modo que se refuerza la seguridad de la clave con la verificación de la identidad del locutor.

Para evitar el robo de las claves es posible que el sistema haga que el usuario repita frases aleatoriamente, de modo que sea difícil de recordar por un impostor. Este tipo de tecnología suele aplicarse en sistemas control de acceso, como por ejemplo banca telefónica o acceso seguro a instalaciones.

•**Independientes de texto**. En este caso el sistema desconoce qué dirá el usuario. Estos son los sistemas más comunes dentro del reconocimiento de locutor.

Podríamos hacer una segunda clasificación atendiendo al modo de operación del sistema:

•**Verificación**. Las aplicaciones típicas de estos sistemas son el control de acceso y autenticación remota (por ejemplo, transacciones telefónicas).

•**Identificación y monitorización**. Entre sus diversas aplicaciones se encuentran la detección de locutores en centralitas y centros telefónicos de atención, o la detección de un locutor en una secuencia de voz (*speaker-spotting*).

•**Sistemas forenses**. Tienen dos posibles aplicaciones. Por un lado pueden ser utilizados para presentar muestras de voz como evidencias en procesos legales. Por otro lado pueden ser utilizados para aportar información en investigaciones forenses.

En este proyecto se experimentará con sistemas independientes de texto de verificación. Dentro de los sistemas de verificación, llegaremos a estudiar el impacto de la calidad al experimentar con bases de datos forenses.

2.4.3 Funcionamiento de un sistema de reconocimiento de locutor.

Todos los SRL se basan en una misma estructura básica para funcionar, independientemente de su aplicación, ya sea identificación, verificación o un sistema forense. Dicha estructura se compone de dos fases:

1. **Fase de entrenamiento**: su finalidad es generar un modelo que sea representativo de la identidad Ij de cada locutor. Para ello se hará uso de distintas muestras del locutor.

2. **Fase de cálculo de la puntuación de similitud**: consiste en comparar una locución Li con un modelo estadístico Mj correspondiente a una identidad Ij. Como resultado de esta comparación se generará una puntuación de similitud *o score*, indicativa de la similitud entre la locución y el modelo.

Dicha puntuación será manejada de una manera dependiendo del tipo de sistema:

•**Verificación**: se establecerá un umbral, por encima del cual se considerará que el modelo y la locución pertenecen al mismo locutor.

•**Identificación**: se llevarán a cabo N comparaciones, donde N es el número de identidades que pueden ser asignadas a un locutor. Con las N puntuaciones obtenidas se tomará una decisión.

•**Sistemas forenses**: en este caso, la locución de test constituye una evidencia y la identidad del acusado viene representada por el modelo Mj. La puntuación obtenida se somete a una transformación para obtener una nueva medida conocida como *Likelihood Ratio* o relación de verosimilitudes, cuya finalidad es apoyar una de las dos posibles hipótesis.

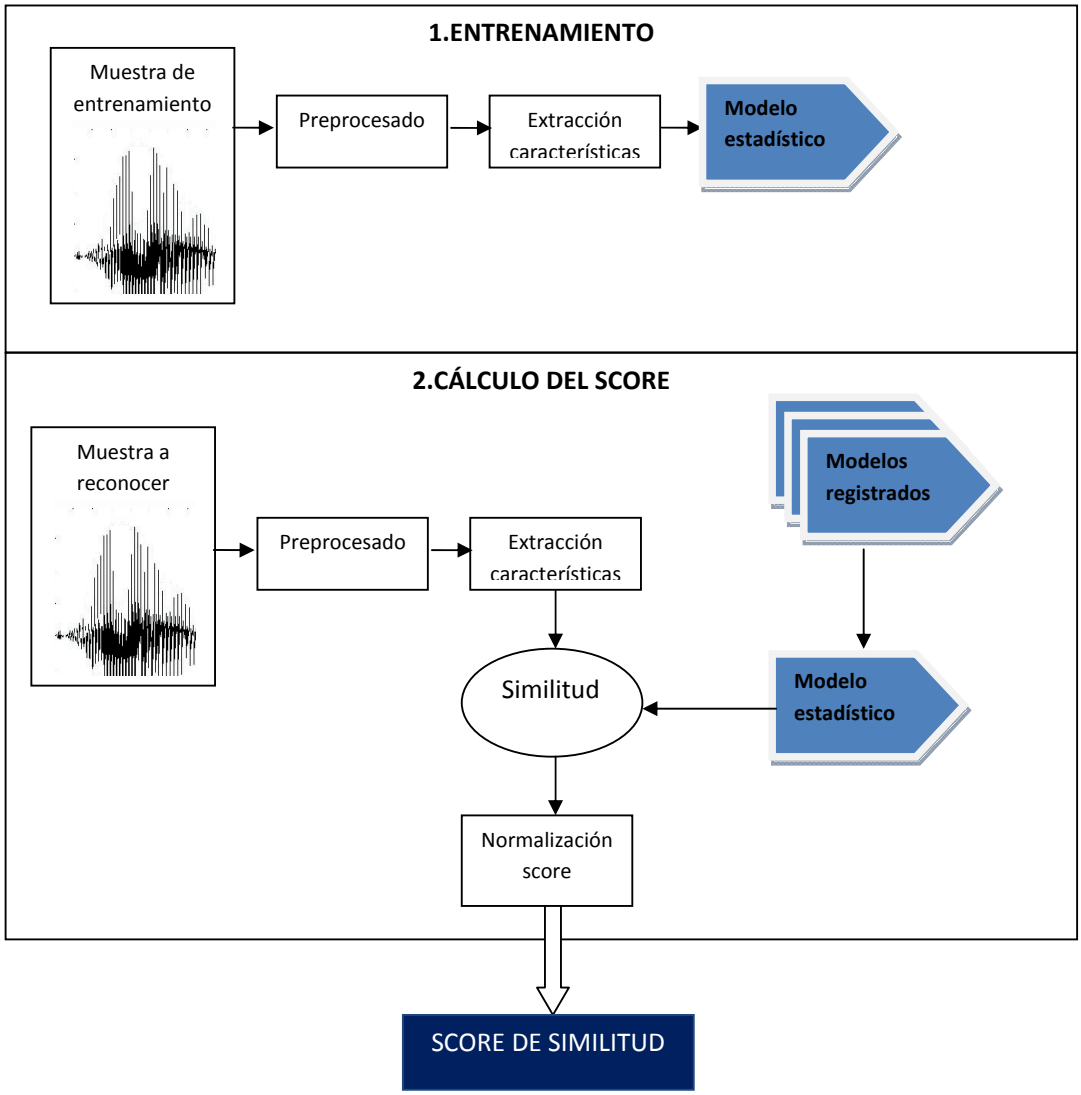


Fig 2.7: *esquema del funcionamiento de un sistema de reconocimiento automático de locutor.*

Existen distintos métodos para medir el rendimiento de un sistema. Los más comunes tratan de medir la capacidad discriminativa. En términos de verificación de usuarios, existen dos tipos de error en los que un sistema puede incurrir: la falsa aceptación y el falso rechazo.

Las curvas DET (*Detection Error Trade-off*) son el método más utilizado, y tratan de representar la capacidad discriminativa de un sistema. Para ello representa la relación entre la falsa aceptación y el falso rechazo, que varía según los valores del umbral de decisión. De este modo, se define el EER (*Equal Error Rate*) como la tasa de error para la cual coinciden la falsa aceptación y el falso rechazo. En el capítulo 4 se explicará su funcionamiento y aplicación con las medidas de calidad.

2.4.4 Tecnologías utilizadas

Haremos una primera clasificación de las técnicas utilizadas atendiendo a los métodos de parametrización del habla.

Extracción de características

Como se ha comentado, se puede extraer información de la señal a distintos niveles. Dicho nivel será el que determine el método de extracción de características y el tipo de modelado estadístico.

•**Sistemas acústicos**. Las medidas acústicas son extraídas directamente del espectro de la señal, por lo que son muy fácilmente accesibles. Existen distintos tipos de parámetros, aunque el procedimiento básico es análogo para todos ellos. Se toman ventanas solapadas de corta duración (decenas de milisegundos) de las cuales se extrae un vector de características.

El procedimiento más común para obtener dicho vector consiste en pasar la señal por un banco de filtros paso banda cuyos centros y espaciado se diseñan acordes con las características de transmisión de la señal para conseguir extraer la mayor cantidad de información posible. El espaciado de los filtros suele ser uniforme por debajo de los 1000 Hz, y no uniforme por encima; de este modo se asemejan al sistema de percepción humano.

Los más utilizados son los MFCC (Mel-Frequency Cepstral Coefficients). Sus filtros son espaciados conforme a la escala Mel (uniforme hasta los 1000 Hz y logarítmico por encima). El inconveniente de estos coeficientes es su sensibilidad a variaciones en el canal o el transductor. Existen técnicas para compensar dichas variaciones: CMS (Cepstral Mean Substraction) , RASTA filtering y feature warping. Se puede consultar más información sobre dichos métodos en [Deller *et al.* 1999].

Otro tipo de parámetros son los LPC (*Linear Predictive Coding*). Son un método comúnmente utilizado en el procesamiento de la señal de audio, que consigue representar la envolvente espectral de la señal mediante un número limitado de coeficientes [Ramos *et. al.,* 2009].

•**Sistemas fonéticos**. Su finalidad es modelar el uso de los fonemas. Están compuestos de dos bloques: el primero se encarga de identificar los fonemas en cada locución. Una vez identificados los fonemas el segundo bloque se encargará de hacer un modelo estadístico del lenguaje del locutor en base al tipo de fonemas y la frecuencia con la cual los utiliza.

•**Sistemas prosódicos**. Se basan fundamentalmente en el análisis de dos factores:

- La frecuencia fundamental (pitch). Viene determinada por la frecuencia de la vibración de las cuerdas vocales, y es la responsable del tono del habla en cada individuo. En [Hess 1983)] se pueden encontrar distintos métodos para el cálculo del "pitch".
- La energía y duración de los sonidos, fáciles de extraer por distintos métodos.

Al igual que los sistemas fonéticos, se componen de dos bloques: el primero se encarga de extraer los parámetros citados. El segundo realiza un modelado estadístico a partir de estos.

Modelado de características

Los sistemas de reconocimiento de locutor más comúnmente utilizados se basan en las características acústicas de la señal. Dentro de este tipo de sistemas, mostramos una segunda clasificación de las técnicas empleadas, atendiendo a los métodos de generación de modelos y cálculo del score de similitud.

Los sistemas basados en Modelos de Mezclas Gausianas **(GMM o *Gaussian Mixture Models)*** consisten en modelar el habla a partir de mezclas de distribuciones gausianas, basándose en que la distribución de los parámetros de tipo perceptivo (LPC y MFCC) se aproxima a la de una mezcla de gausianas. De modo que a cada usuario o población de interés le corresponderá un modelo λ , cuya función de densidad de probabilidad vendría dada por [Bimbot *et. al*.2004]:

$$f(x|\lambda) = \sum_{i=1}^{M} w_i f_i(x)$$

donde la variable independiente x representa un vector de características, w_i es el peso de cada una de las M gausianas $f_i(x)$, que a su vez se descomponen en:

$$f_i(x) = \frac{1}{(2\pi)^{d/2}|\sum i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T(\sum i)^{-1}(x-\mu_i)}$$

siendo μ_i y $\sum i$ las medias y la matriz de covarianzas de las gausianas y d el número de dimensiones del modelo.

Para generar cualquier score de similitud se utilizarán dos modelos estadísticos. Por un lado se tendrá un modelo de habla universal (UBM o Universal Background Model) que representará el tipo de habla genérica del entorno estudiado. Dicho modelo generalmente se entrena con bases de datos de grandes dimensiones que representen la población bajo estudio. Por otro lado, se tendrá el modelo del locutor cuya identidad se quiere verificar, el cual se obtiene adaptando el UBM a los parámetros extraídos de las locuciones de entrenamiento de dicho locutor.

De este modo, sea $\mathrm{O} = \{\mathrm{O}_1, \mathrm{O}_2 \dots \mathrm{O}_\mathrm{N}\}$ la secuencia de observaciones de un vector de dimensión extraído de un segmento de voz, y λ_t el modelo generado usando la información del usuario, definiremos la medida de similitud entre ambos, puntación, como $\mathrm{S}(\mathrm{O}, \lambda_\mathrm{t})$, y será calculada evaluando las funciones de densidad de probabilidad de ambos modelos comentados, para el vector de características extraído del segmento de voz:

$$S(O, \lambda_t) = log\, f(O|\lambda_t) - \, logf(O|\lambda_{UBM})$$

Donde $\mathrm{f}(\mathrm{O}|\lambda_\mathrm{t})$ y $\mathrm{f}(\mathrm{O}|\lambda_{\mathrm{UBM}})$ son las funciones de densidad de probabilidad evaluadas para el vector O.

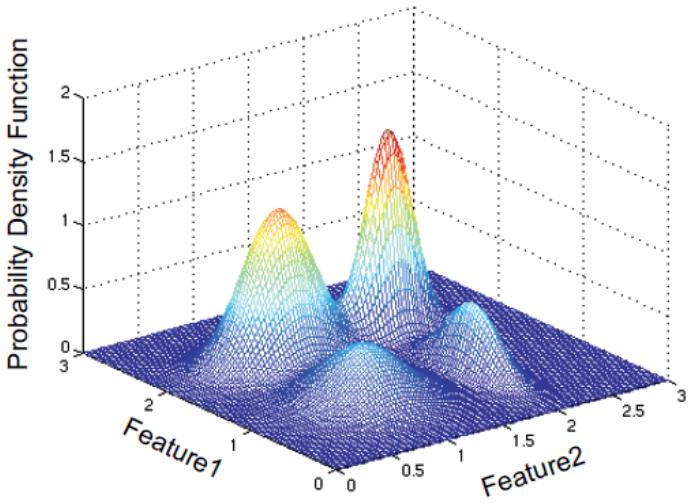


Fig 2.8*: ejemplo de GMM. Figura adaptada de [Ramos, 2007]*

Las **Máquinas de Vectores Soporte (SVM o *Support Vector Machines*)** son una técnica de aprendizaje discriminativo que en los últimos años ha mostrado un rendimiento equivalente al de los GMM.

Su objetivo es establecer la frontera que separa a dos clases.
Partiendo de dos conjuntos de vectores asociados a dos clases (p.ej. el habla del usuario y el habla universal), el objetivo de los SVMs será establecer un hiperplano de separación de las dos clases.

Los datos con los que entrenaremos el sistema serán una serie de vectores etiquetados, de la forma: $\{\overrightarrow{x_i}, y_i\}$ $i = 1, \dots, l$, donde:

$\overrightarrow{x_i} \in R^d$ es el vector de observaciones en un espacio de dimensión d.
$y_i \in \{-1,1\}$ representa la etiqueta de la clase a la que pertenece cada vector.

El problema consistirá en asignar cada vector a su clase correspondiente, 1 ó -1, para ello se calculará un hiperplano de separación que divida el espacio R^d en dos regiones. Una vez establecido dicho hiperplano, las muestras que caigan en una región pertenecerán a clase -1 y las que caigan en la otra a la clase 1.

El cálculo de este hiperplano supone que los vectores de ambas clases son linealmente separables. Esto en la realidad no es así; por ello se tienen que definir funciones de mapeo o kernels que transformen los vectores en el espacio original $\mathrm{R^d}$ a un espacio de mayores dimensiones $\mathrm{R^{d'}}$, donde $\mathrm{d'} > d^{\mathrm{d}}$. Existen diversos tipos de kernels (lineales, polinómicos, radiales, etc.). El kernel GLDS (*Generalized Linear Discriminative Sequence*) es el más comúnmente utilizado, y será el empleado en nuestro estudio [Campbell 2002].

Una vez establecido el hiperplano de separación entre las dos clases, el score de similitud consiste en la distancia desde el vector correspondiente al fichero de test, al hiperplano entrenado como modelo del locutor.

Una variante de los sistemas SVM [Campbell 2006] denominada **SVM Supervector** consiste en entrenar un modelo GMM para cada locución que interviene en una comparación (tanto de testeo como de entrenamiento) para después representar dicho modelo como un vector de un número elevado de dimensiones llamado supervector, de modo que pueda ser clasificado utilizando un sistema SVM. El proceso de cálculo del score en este tipo de sistemas se resume en estos pasos [Mateos, 2007]:

1. Se entrena un modelo de GMM para cada locución que intervenga en el experimento.

2. Cada modelo estará compuesto de una mezcla de gausianas. El supervector se formará extrayendo y concatenando las medias de cada una de las gausianas ponderaras por su peso y su matriz de covarianza. Este vector tendrá una dimensión $m * d$, donde m es el número de gausianas y d es el número de dimensiones.

3. La discriminación entre supervectores correspondientes a locutores impostores, NonTarget, y locutores usuarios, Target, la llevaremos a cabo mediante un sistema SVM, que es básicamente un clasificador binario.

4. La puntuación de un enfrentamiento entre un modelo y un fichero de test se obtendrá de la misma forma que en un SVM clásico. Enfrentando el modelo (hiperplano) del locutor correspondiente con el supervector de la locución de test y calculando su distancia a dicho hiperplano.

Técnicas de compensación

En este apartado se presentarán distintas técnicas, que llevadas a cabo a nivel de parámetros, tratarán de compensar la influencia del ruido y otros efectos perturbadores en la señal.

• **Normalización por media cepstral** (CMN o *Cepstral Mean Normalization*)
Esta técnica es una de las más populares desarrolladas en el campo de la normalización de canal. Consiste básicamente en restar a los vectores de parámetros la media de dichos vectores estimada a lo largo de todo el fichero, bajo la hipótesis de que el canal es un elemento de variación lineal en el dominio cepstral y que por tanto su contribución principal es a la media de los vectores cepstrales. En [Furui, 1981; Garcia y Mammone, 1999] puede encontrarse una descripción más amplia de la teoría cepstral y de CMN.

• **RASTA filtering**
El filtrado RASTA [Hermansky y Morgan, 1994], al igual que la normalización CMN, va orientado a eliminar las distorsiones introducidas por el canal. La complejidad de esta técnica es superior a la de CMN, haciendo que el espectro de la señal de voz resultante dependa de instantes pasados y realzando las transiciones espectrales.

• **Feature Warping y Feature Mapping**
La técnica Feature Mapping [Reynolds, 2003b] es una técnica de normalización orientada a los datos que presenta unos resultados mejores que Feature Warping [Pelecanos y Sridharan, 2001]. En feature warping el objetivo era conseguir una distribución final gaussiana de media nula y varianza unidad (técnicas de transformación de histograma), tratando cada dimensión del vector de parámetros por
separado. En *Feature Mapping*, por su parte, se tiene en cuenta la correlación entre dimensiones a la hora de realizar la normalización.

• **Factor analysis.**

Es una técnica de compensación desarrollada en los últimos años y que ha mostrado excelentes resultados [Kenny, 2004]. Consiste en modelar las direcciones de máxima variabilidad interlocutor e intralocutor (debidas a la variabilidad intersesión) de las características extraídas del habla. A partir de esta información se intentan compensar aquellas variaciones relacionadas con la variabilidad intralocutor y mantener las interlocutor.

3. CALIDAD

3.1 Calidad en biometría

En este capítulo se presentan los aspectos más importantes de la calidad en reconocimiento biométrico basándonos en otros trabajos del estado del arte.

3.1.1 Definiciones calidad

Al igual que sucede con las personas, hay dos factores que determinan la precisión de una comparación: el poder discriminativo del sistema y la cantidad de información disponible. Dentro de este último, existen otros dos factores determinantes: la cantidad de muestras tomadas y su fidelidad o similitud con la fuente original. Y a su vez esta fidelidad vendrá determinada por otros dos factores: el comportamiento del individuo y las distorsiones generadas durante su captura, transmisión o almacenamiento (compresión).

[M1/05-0306] es un borrador de un estándar de calidad de muestras biométricas. En él se define la calidad conforme a tres criterios. Dos de ellos son la fidelidad y el carácter. Este último se define como todas aquellas características físicas o del comportamiento del individuo que determinan la probabilidad de poder distinguirlo de otro (distintividad, universalidad, actitud durante la captura, etc.). Dentro de la fidelidad, se distinguen tres posibles procesos causantes de degradaciones en las muestras: adquisición de la muestra, procesado y extracción de características.

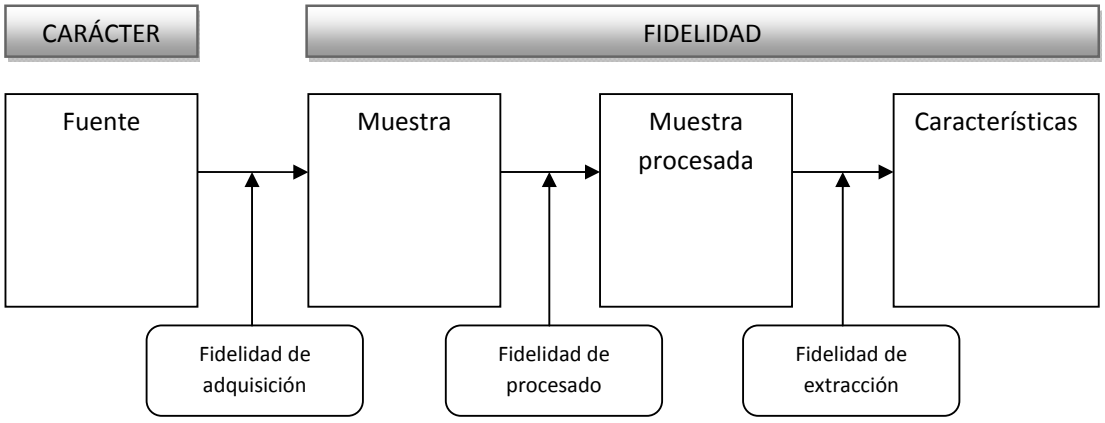


Fig. 3.1: *esquema de los factores influyentes en la calidad de las muestras biométricas.*

El tercer criterio que se define en [M1/05-0306] es la utilidad, la cual se basa en predecir el impacto que tendría la calidad de una muestra biométrica en un sistema. De este modo, cuanto mayor sea la calidad de la muestra, menor debería ser la probabilidad de que el sistema cometa un error clasificándola. Bajo este criterio, tanto factores del carácter (atribuibles al individuo) como de la fidelidad influyen en la calidad de la muestra biométrica.

En el caso que nos ocupa, la única manera de medir si los algoritmos de medición de calidad que utilizamos son correctos, es observar su relación con el rendimiento de los sistemas. Por lo tanto las medidas de calidad estarán definidas bajo el criterio de utilidad [Alonso *et al.* 2008;Grother *et al.* 2007].

3.1.2 Factores degradantes de la calidad

•Factores relacionados con los sujetos (carácter)

Como se ha comentado en el apartado anterior, las características físicas y del comportamiento de un sujeto pueden afectar a la calidad de la muestra. Las que conciernen a los individuos se pueden clasificar en las siguientes [Hicklin, 2006]:

•Información limitada por las **características del sujeto**. Un buen ejemplo es el reconocimiento de cara en niños. Este es más difícil que en adultos debido a los cambios de la piel y estructura facial que todavía no han sufrido. Del mismo modo una huella dactilar podría ser capturada con una nitidez excelente, pero ser difícilmente distinguible por el hecho de contener muy pocas minucias.

• **Cambios en las características**. Pueden producirse cambios en un rasgo por distintos motivos: golpes, operaciones estéticas, crecimiento de bello, enfermedades, alteraciones emocionales o enfermedades. Un claro ejemplo en el caso de la voz, sería una afonía o un constipado, que hacen cambiar las características de la voz.

• **Comportamiento del sujeto**. Hay varias maneras por las que esto puede afectar a la calidad de una muestra: la cooperación del sujeto, el grado en el que está habituado, su conocimiento sobre la captura de los rasgos y el estado emocional son los más importantes. En sistemas forenses de reconocimiento de locutor este es un factor especialmente importante. Es común que el sujeto se encuentre en un estado emocional alterado, o que intente modificar su voz para evitar ser reconocido.

• **Fraude**. Podemos distinguir dos tipos de fraude:

- **Evasión**: es muy común en casos forenses que el sujeto intente modificar u ocultar algún rasgo para evitar así ser identificado. En casos de evidencias de voz, es habitual que el locutor intente distorsionar su voz de alguna manera.
- **"Spoofing":** consiste en intentar culpar a otro sujeto mediante una muestra biométrica del mismo.

(a) (b)

Fig 3.2: (a): imagen de huella dactilar utilizada en caso de evasión. (b): mismo rostro modificado gráficamente. Figura adaptada de [Hicklin, 2006]

Factores relacionados con la adquisición de datos

Existen dos factores degradantes relacionados con la adquisición de datos:

• **Dispositivos de adquisición de datos**. Distintos dispositivos de captura extraen distinta información con distintos grados de fidelidad. Un claro ejemplo son las huellas dactilares. En [Alonso *et. al.*, 2008] se experimentan con medidas de calidad en tres tipos de sensores: capacitivos, ópticos y térmicos. En este proyecto también se experimentará con dos tipos de captura del audio: datos telefónicos y microfónicos.

• Los **procesos de adquisición de datos**. Factores tales como la supervisión del sujeto durante una adquisición, el número de imágenes tomadas (en el caso de huella o iris, por ejemplo), el tiempo de habla registrado (en el caso de voz) o las condiciones ambientales de la adquisición (temperatura, humedad, superficies reflectantes, salas con reverberaciones, etc.) influyen en la calidad de la muestra biométrica.

Factores relacionados con el procesado y la compresión de los datos

En sistemas de reconocimiento de locutor, el ejemplo más claro de procesamiento de la señal es la transmisión de la misma a través de la red telefónica. La segmentación (en caso de imagen) o el uso de detectores de actividad de voz (VAD) también pueden producir pérdida de información.

Por otro lado, a menudo se almacenan las muestras biométricas en formatos comprimidos. Esta compresión puede dar lugar a pérdida de información.

Factores relacionados con la extracción de características

Diferentes métodos de extracción de características extraen diferente información de las muestras. Dicha información vendrá caracterizada por un grado de fidelidad. Como ejemplo, podemos suponer el caso de la extracción de parámetros de voz, en el que un número inapropiado de parámetros puede producir una pobre representación de la muestra.

3.1.2 Consecuencias de variabilidad de la calidad.

Un descenso de la calidad es equivalente a una reducción de información de la muestra biométrica. Al disponer de menos información es menos probable que el sistema consiga una clasificación correcta, aumentando las tasas de error, ya sea a corto o a largo plazo. Esta es una consecuencia inmediata si se reduce significativamente la calidad de la muestra. Sin embargo, es posible que las consecuencias sólo sean detectadas a largo plazo [Hicklin, 2006]. Un ejemplo es la robustez que presenta un sistema al envejecimiento de un usuario. Cuanto mejor sea la calidad del modelo que representa a dicho usuario, afectará en menor medida las posibles variaciones de los rasgos del individuo.

3.1.3 Aplicaciones de la calidad.

• **Recapturación** [Hicklin, 2006]. Consiste en medir la calidad de las muestras y rechazar aquellas que no superen un determinado umbral. Se vuelve a capturar el rasgo biométrico tantas veces como sea necesario para asegurar un umbral mínimo de calidad. Resulta especialmente útil en sistemas de acceso on-line, y para la recolección de muestras para generar modelos.

• **Generación de modelos**. Las muestras a partir de las cuales se forman los modelos habitualmente están sujetas a una calidad variable, ya sea entre las distintas muestras, o para una misma muestra en distintas zonas de la misma. La medición de la calidad permite reducir el efecto de aquellas muestras (y porciones de muestras) con una calidad peor. Esto se consigue otorgando un menor peso en el modelo a toda información de menor calidad, ya que cuanto menor es esta, menor es la información que contiene sobre un sujeto y mayor es la probabilidad de inducir errores de clasificación.

• **Comparación**. Se basa en la misma idea que la generación de modelos, pero aplicado a la calidad del score global. Otorgando un peso mayor a aquellas secciones de la muestra biométrica con mejor calidad, es de esperar que mejoren los resultados si la medida de calidad es útil.

• **Fusión dependiente de calidad [**Alonso, 2008b**;**Fierrez *et al*. 2004**]**. Distintos sistemas se ven afectados en diferente medida por los distintos factores de degradación. Aprovechando esto, se puede hacer una fusión dependiente de la calidad, otorgando un mayor peso a los sistemas más robustos ante un descenso de la calidad.

• **Calibración dependiente de calidad**. Consiste clasificar un grupo de scores en función de su calidad, para después aplicar transformaciones diferentes a cada uno de los grupos. Si la calidad consigue determinar grupos de scores que puntúan en rangos diferentes, la calibración colocará los nuevos scores en rangos similares, de modo que mejore el rendimiento del sistema. La regresión logística [Bishop, 2006] es un método de calibración que ha demostrado eficacia con este tipo de compensación. En [Ferrer, 2008] se utiliza información auxiliar (número de fonemas, índice de "no natividad" de un locutor) sobre la señal de audio para aplicar regresión logística a los scores.

• **Normalización dependiente de calidad.** Consiste en aplicar diferentes parámetros de normalización en función de la calidad de la muestra. En [Alonso, 2008b] se utiliza este método.

3.2 Calidad en voz

3.2.1 Antecedentes

La estimación de la calidad de la voz es un campo ampliamente estudiado desde hace años con el fin de mejorar la calidad del servicio en redes de comunicaciones telefónicas. Los primeros métodos para estudiar la calidad de la voz se basaban en observar la calificación otorgada por personas a una serie de locuciones, lo que se conoce como calidad subjetiva. Estos métodos demostraron ser eficaces, y llegaron a establecerse una serie de recomendaciones que definían los protocolos experimentales para llevar a cabo dichas mediciones [ITU-P.800]. Con el paso del tiempo estos métodos se quedaron obsoletos por dos razones: son muy costosos en tiempo y dinero, y por otro lado las nuevas tecnologías de voz (principalmente comunicaciones móviles y voz por IP) hacen necesario monitorizar la calidad constantemente y en diversos puntos de la red de comunicaciones.

Los esfuerzos llevados a cabo en este campo han dado como fruto distintas herramientas, estándares y recomendaciones, cuya documentación ha aportado amplios conocimientos sobre estimación de calidad para aplicarlos en el presente proyecto [Malfait *et al.* 2004; Grancharov, 2007].

3.2.2 Tipos de medidas de calidad

En esta sección se lleva a cabo una clasificación de las principales medidas de calidad utilizadas en voz hasta este momento, y cuya aplicación a SRL es posible. Es importante tener en cuenta que no todas las medidas han sido utilizadas en sistemas de reconocimiento de locutor, pues pueden haber sido estudiadas con otros propósitos como la estimación objetiva de calidad subjetiva.

Se han definido los siguientes grupos en base a los procedimientos utilizados en la estimación de la calidad:

- **Medidas de estimación de la calidad subjetiva.**

Su propósito es estimar la calidad de la señal de voz aproximándose lo más posible al modelo psicoacústico humano. Suelen ser medidas enfocadas a la monitorización de la calidad del servicio en redes de comunicaciones.

La más representativa, es la recomendación de la ITU, ITU-P.563 [Malfait *et al*. 2004], de la cuál hablaremos en el siguiente apartado, y que será utilizada en nuestro estudio.

- **Análisis del ruido**

Su finalidad es estimar el nivel de ruido de una locución. Existen distintos métodos para esto, y en realidad, la mayoría de las medidas de calidad que se utilizan están ligadas al nivel de ruido. La más común es la relación señal a ruido (SNR). Para su cálculo existen distintos métodos, el más común de todos y el más simple consiste en calcular la energía del ruido en los silencios de las grabaciones y la energía de la voz en zonas de habla, para después calcular la energía media de cada uno de ellos. Este método tiene algunos inconvenientes, a saber:

- Imprecisión en locuciones con alta tasa de actividad de voz: ya que es posible que la locución no contenga suficiente tiempo de silencio como para estimar correctamente el nivel de ruido.
- No se tienen en cuenta ruidos presentes en zonas con voz, que pueden ser causadas por micrófonos, amplificadores, codificaciones, etc.
- Depende de la disponibilidad de los silencios para ser calculada. Como se podrá comprobar en la sección de experimentos, no siempre se dispone de los silencios.

• Análisis del tracto vocal

En este tipo de estimaciones se pretenden identificar distorsiones en las locuciones basándose en modelos del tracto vocal asociados a dichas locuciones. Dichos modelos se realizan asociando a cada cavidad del tracto vocal un tubo de diámetro X y longitud Y, que varían con el tiempo.

Para ello se definen una serie de reglas o baremos para identificar violaciones de lo que se considera el habla normal. Por ejemplo, suponiendo que disponemos una representación del tracto vocal como la descrita más arriba, una posible violación sería un incremento notable del diámetro de una cavidad en un pequeño espacio de tiempo. En la siguiente imagen se puede observar un ejemplo, extraído de la documentación de la ITU-563 [ITU-P563].

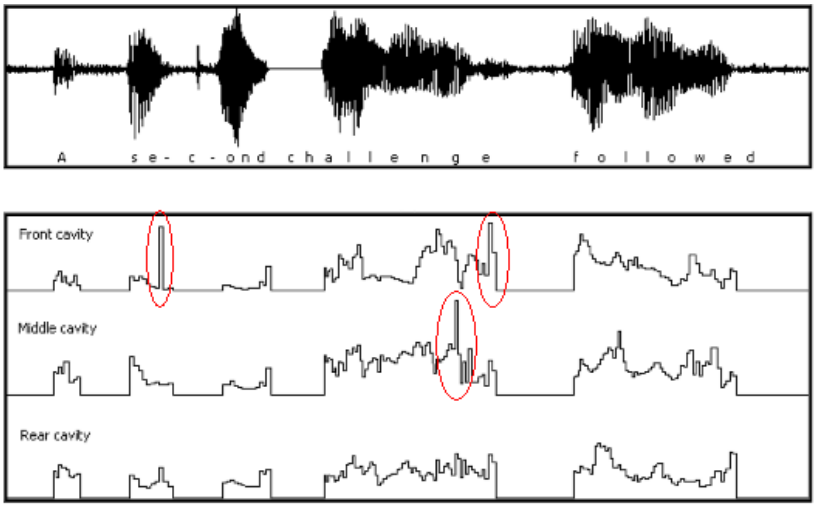


Fig 3.3: ejemplo de modelo de tracto vocal para estimación de la calidad de la voz. Imagnen extraída de [P.563]

Este método se ha encontrado en otros estudios del estado del arte [ITU-P563; Gray *et al*. 2000]. El primero corresponde a la recomendación P563, en la cual se extraen un total de 14 parámetros con este método para estimar la degradación de la voz. El segundo corresponde a un estudio enfocado a la estimación objetiva de la calidad subjetiva de la señal de voz, utilizando únicamente este tipo de métodos.

• **Estadísticas de la voz.**

Este tipo de medidas consisten en obtener diferentes tipos de medidas estadísticas de la señal de voz que sean indicativas de la degradación de la misma.

En [ITU-P563] se utilizan dos medidas estadísticas llamadas skewness y kurtosis las cuales se aplican a los parámetros MFCC y LPC de la señal.

En [Richiardi y Drygaijlo, 2007] se utilizan también la skewness y la kurtosis aplicadas a la distribución en el dominio temporal de toda la locución.

En [Gª Romero *et. al.,* 2005] se propone una medida de calidad basada en las variaciones de la frecuencia fundamental de la voz.

• **Basadas en modelos estadísticos**

A este último grupo pertenece una de las medidas utilizadas en este proyecto. No se ha encontrado en el estado del arte ninguna medida similar, por lo que constituye una contribución del presente proyecto. Básicamente consiste en aprovechar modelos estadísticos de habla poco degradada para determinar la calidad. Se verá con más detalle en los tipos de medidas (siguiente apartado).

3.2.3 Marco teórico para la medición de calidad

Basado en trabajos previos en la materia [Gª Romero *et al.* 2005;Grother *et al.* 2007], y en las recomendaciones del NIST [M1/05-0306], se ha definido medida de calidad como una magnitud escalar que predice el rendimiento de un sistema biométrico. Dicha magnitud estará limitada entre los valores 0 y 1, donde el 0 se corresponde con la peor calidad que puede ser atribuida, y el 1 a la mejor.

Generalmente las medidas están basadas en parámetros que no están dentro del rango comentado (p.ej. la SNR), por lo que hay que definir una función de mapeo $Q(x)$ que a cada valor de un parámetro x le asigne un valor entre 0 y 1.

Cuando se conoce el impacto que tiene un parámetro en un SRL es fácil conocer cómo aplicar la función de mapeo. Por ejemplo, suponiendo que queremos establecer la relación señal a ruido (SNR) como medida de calidad, la función de mapeo debería asignar un 0 a los valores más bajos posibles de SNR, y un 1 al más alto posible. Sin embargo, en este estudio se tratan parámetros de los cuales se desconoce su impacto en SRLs, por lo que se identificó la necesidad de hacer un estudio previo sobre cada uno de estos parámetros, a los que llamaremos indicadores de degradación (habitualmente serán referenciados por su acrónimo ID).

El estudio de los indicadores de degradación (IDs) permitirá conocer en qué medida son útiles, y proporcionarán información suficiente para diseñar la función de mapeo correspondiente.

Dado el enfoque de utilidad adoptado para comprobar la validez de las medidas de calidad, será necesario definir la calidad de una comparación o calidad del score, que será utilizada para tratar las comparaciones de acuerdo a sus calidades. Dicha calidad del score se define como:

$$Q = \sqrt{Q_{test} \cdot Q_{train}}$$

donde Q_{test} y Q_{train} son respectivamente las calidades asociadas a las locuciones de testeo y entrenamiento.

3.3 Indicadores de degradación estudiados

Las fuentes utilizadas para escoger los indicadores de degradación provienen tanto de estudios sobre calidad subjetiva de la voz, como de los escasos estudios de calidad en SRLs ya existentes, y también otros que se proponen en este trabajo, que no se han encontrado en el estado del arte. El hecho de que en la literatura este tipo de estudios sea escaso le da más relevancia e impacto científicos a este PFC.

A continuación se explican los indicadores de degradación utilizados.

3.3.1 ITU-P.563

Es un método de evaluación de calidad perceptiva. La calidad percibida es en esencia subjetiva [Grancharov *et al.* 2007], es decir, varía de individuo a individuo. P-563 es el estándar de la ITU (*International Telecommunications Union*) para predecir objetivamente la calidad subjetiva.

Probablemente esta calidad perceptiva o subjetiva no determine con precisión la calidad para un SRL, ya que no está diseñado para tal fin, pero sin ninguna duda es un indicador de la degradación de la señal, por lo que conviene estudiar su efecto este tipo de sistemas, máxime cuando los SRLs utilizan frecuentemente enfoques perceptuales para extraer características de la voz (MFCC, PLP).

Esta recomendación surgió por la necesidad de monitorizar la calidad de servicio de las redes telefónicas en tiempo real. Para cubrir dicha necesidad la ITU abrió un concurso que dio como fruto dicha recomendación. Actualmente uno de los métodos más populares en la medición de calidad subjetiva de voz proveniente de habla telefónica. Tiene en cuenta la mayoría de factores degradantes de la señal en redes de comunicaciones modernas tales como ruido, ecos y reverberaciones, *jitter*, pérdida de paquetes, efecto local etc.

La ITU proporciona una implementación de su recomendación, que incluye una amplia documentación en la materia. El estudio de esta documentación permitió profundizar en la materia, de modo que puede considerársela una buena referencia para cualquier estudio sobre estimación de calidad en voz, ya sea para sistemas de reconocimiento u otras aplicaciones [ITU-T P563].

La herramienta, de una gran complejidad, calcula un total de 51 parámetros con los cuáles estima el tipo de degradación dominante en la señal, al igual que calcula una puntuación sobre la escala MOS (*Mean Opinion Score*) que será representativa de la calidad de la señal.

Para calcular el tipo de degradación dominante (p.ej: ruido, desnaturalización de la voz, cortes, etc.) se analizan 8 de esos parámetros, cada uno de los cuales guarda un alto grado de correlación con un factor de degradación determinado. Esta idea sirvió para determinar qué indicadores de degradación estudiar para este PFC, de modo que se seleccionaron 5 de los parámetros (que serán explicados en este mismo apartado) atendiendo a su importancia en el peso final sobre el MOS, y a cuán asequible era su implementación.

Esta herramienta, a pesar de ser muy completa, cuenta con una serie de limitaciones. Los archivos de audio deben tener una duración de entre 3 y 20 segundos. Sin embargo, en las bases de datos con las que se experimenta en este trabajo, todos los archivos tienen duraciones de entre 2 y 5 minutos .

Esto se solucionó de la siguiente manera:

1. Se dividió la locución en fragmentos de entre 3 y 20 segundos.

2. Se evaluó la calidad de cada uno de ellos mediante la ITU-T P.563.

3. Se hizo el promedio de todos ellos con un peso proporcional a su duración.

Se llevaron a cabo varias comprobaciones para ver si coincidía la calidad de un archivo con la media ponderada de los fragmentos, y no fue así. Sin embargo, como se mostrará más adelante, sí guarda cierta relación con la calidad de la señal para un SRL.

3.3.2 SNR

Las siguientes medidas tratan de estimar la relación señal a ruido (SNR). Para ello, se estima la potencia del ruido de dos formas diferentes: a partir de los silencios de las locuciones y a partir de la señal limpia reconstruida mediante filtrado adaptativo. Podría hacerse una estimación más precisa, ya que existen ruidos de tipo multiplicativo que sólo se encuentran en las secciones de voz, pero la estimación de estos últimos requiere de algoritmos más complejos y costosos en tiempo [ITU-P563], por lo que sólo se estimarán los primeros.

A continuación se explican los métodos mencionados, que calculan el ruido de deiferente manera:

- **De los silencios de la señal (“SNR”)**

Haciendo uso de un detector de actividad de voz (DAV) se identifican las tramas de voz activas y las no activas (silencios). A continuación se calcula la energía de cada una de ellas para después calcular la energía media de la voz y los silencios. Por último se obtiene la SNR como

$$SNR = 10 * \log\left(\frac{E_{voz}}{E_{sil}}\right)$$

Siendo E_{voz} y E_{sil} la energía media de las zonas de voz y silencio respectivamente.

- **De la misma señal filtrada (“SNR-WIENER”)**

El filtro de Wiener, es un método que ha demostrado gran eficacia en la reducción de ruido [Chen *et al.*]. Es un filtro adaptativo cuyas especificaciones varían en función de la distribución espectral del ruido, que se estima en los silencios.

El cálculo de este indicador de degradación consiste en lo siguiente: a partir de la señal filtrada, la obtención de la señal de ruido se reduce a substraer las señales original y filtrada, de modo que la señal resultante debería ser la señal de ruido eliminado por el filtro. Una vez obtenida la señal de ruido, se calculan las potencias de señal y de ruido y se aplica la fórmula citada anteriormente.

Para sustraer las señales es necesario alinearlas, ya que al aplicar un filtro digital puede verse reducida la longitud de la señal filtrada debido a las colas del mismo.

La idea de aplicar este método surgió a raíz de observar que al aplicar un filtrado de Wiener a las locuciones, los resultados de ciertos SRL mejoraban notablemente su rendimiento. No se ha encontrado otro estudio que utilice este indicador de degradación, constituyendo otra contribución de este trabajo.

3.3.3 Medidas estadísticas de la voz

La *skewness* y la *kurtosis* son dos medidas estadísticas que han sido utilizadas, entre otras aplicaciones, para determinar en qué medida se aproxima una distribución a una gaussiana. La kurtosis mide cómo de picuda es la distribución, mientras que el sesgo mide la simetría de la misma [Richiardi y Drygajlo, 2007].

El sesgo se calcula con el momento estadístico de tercer orden:

$$s = \frac{1}{P}\sum_{p=1}^{P}\left(\frac{a_p - \frac{1}{P}\sum_{p=1}^{P} a_p}{\sigma}\right)^3$$

donde a_p serían los valores cuya distribución se quiere evaluar y σ la desviación típica. Cómo se ha comentado, mide el grado de simetría de una distribución. Será negativo si la cola izquierda es más larga y positivo en caso contrario.

La kurtosis se calcula con el momento estadístico de cuarto orden de la distribución.

$$k = \frac{1}{P}\sum_{p=1}^{P}\left(\frac{a_p - \frac{1}{P}\sum_{p=1}^{P} a_p}{\sigma}\right)^4 - 3$$

Mide la relación entre el pico y las colas de la distribución. Está normalizada de modo que una distribución gaussiana tiene una kurtosis igual a 0.

Estas dos medidas estadísticas aplicadas a diferentes tipos de información, proporcionan diferentes indicadores de degradación. En nuestro caso se implementaron las que se detallan a continuación:

• En el dominio del tiempo (“Kurtosis Global” y “Skewness Global”)

Este método fue utilizado en [Richiardi y Drygaijlo, 2007]. Se trata de calcular ambas medidas sobre el archivo de voz entero, incluyendo los silencios, de modo que el vector a_p de la fórmula correspondería a los valores de la señal en el dominio del tiempo. En una locución poco ruidosa, se espera que gran parte de las muestras estén en el valor 0 o muy próximas al mismo, de modo que su distribución debería presentar un pico importante en el valor 0. Además si se trata de voz poco degradada, también debería mostrar una distribución simétrica. Por último, la distorsión en las muestras de audio debería también alterar la kurtosis de las mismas.

• Kurtosis calculada a partir del histograma de muestras de audio ("Kbins").

La kurtosis presenta el inconveniente de ser muy sensible a "outliers" (escasos puntos muy alejados de la distribución que presentan el resto de los puntos bajo análisis). En [Richiardi y Drygaijlo, 2007] se propone una medida alternativa a la kurtosis que consiste en representar el histograma de la locución en 100 intervalos y dividir la amplitud del intervalo central entre la suma del resto de las amplitudes. De esta manera se aproxima cómo de picuda es la distribución.

$$Kb = \frac{Nc}{Nt - Nc}$$

Donde Nc sería la frecuencia correspondiente al intervalo central y Nt la cantidad de puntos de la señal.

En el citado estudio se muestra que en las condiciones experimentales presentadas esta medida guarda una estrecha relación con la relación señal a ruido.

En la siguiente figura se muestran los histogramas de las muestras temporales de dos locuciones de la base de datos SRE 2006 con niveles de ruido muy diferentes.

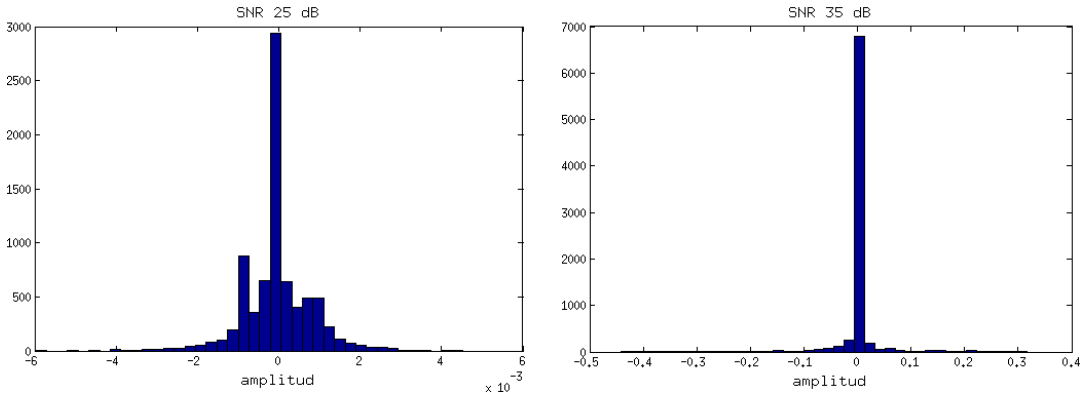


Fig 3.4: *ejemplo de distribuciones para dos locuciones con SNRs distitnas.*

En este caso se observa claramente que en la locución más ruidosa hay una proporción menor de valores con amplitud 0.

• En el dominio del tiempo sobre las tramas activas de voz ("KURTOSIS LOCAL" y "SKEWNESS LOCAL").

En este caso se aplican la kurtosis y skewness sobre la distribución en el dominio temporal, escogiendo tramas de 20 ms activas (sin silencios). Se aplica la fórmula a cada una de las N tramas, y se calcula la media aritmética. Se hace uso de un detector de actividad de voz para separar tramas activas de tramas no activas.

Este indicador no se ha encontrado en ningún otro estudio del estado del arte, siendo una aportación de este trabajo.

• Sobre los coeficientes LPC ("KURTOSIS LPC" y "SKEWNESS LPC")

Estas medidas son dos de las escogidas de la documentación de la ITU-P.563 [ITU-P563]. El procedimiento para calcularlas es el siguiente:

1. Se segmenta el archivo de voz con el DAV, y se toman N tramas de voz de 20 ms.
2. Se calculan 21 coeficientes LPC de cada trama de voz.
3. Se calculan la kurtosis y skewness sobre dichos vectores de coeficientes.
4. Se hace la media aritmética de los N valores de kurtosis o skewness.

• Sobre los coeficientes Cepstrales("KURTOSIS CEPSTRAL" y "SKEWNESS CEPSTRAL")

También fueron extraídas de la recomendación P563 [REF]. El procedimiento es análogo al anterior, pero en este caso se calculan 21 coeficientes MFCC y se calcula su kurtosis y su skewness.

3.3.4 Contribución del PFC: similitud a un modelo de habla universal (UBML)

Esta medida es una contribución original del presente proyecto fin de carrera, no ha sido encontrada en ningún estudio sobre calidad subjetiva ni objetiva. Trata de aproximar la similitud entre una locución y el modelo universal utilizado para generar el modelo estadístico de un locutor (sección 2.4.3).

Su obtención es inmediata si se utiliza un sistema de reconocimiento de locutor basado en GMM, ya que para calcular un score de similitud es necesario calcular la verosimilitud entre el UBM del modelo y la locución de testeo.

Si recordamos la fórmula para la extracción del score de similitud, venía dada por el logaritmo de un cociente de similitudes:

$$S(O|\lambda_t) = \log \frac{f(O|\lambda_t)}{f(O|\lambda_{UBM})}$$

donde el numerador corresponde a la similitud de la locución con el modelo estadístico del locutor. El ID propuesto corresponde al denominador, $f(O|\lambda_{UBM})$, que es la verosimilitud del habla de testeo frente a la función de densidad de probabilidad del modelo universal. Esta magnitud está relacionada con la similitud de la locución con respecto al modelo de habla universal.

Este ID se basa en la idea de que si un modelo universal está entrenado con un tipo de habla, una locución de un tipo de habla diferente va a funcionar peor porque el UBM no es representativo, y por tanto se le debe asociar una calidad baja. Así, este ID da una idea del

desajuste de las locuciones que se van a utilizar con el sistema con respecto del habla utilizada para entrenar el mismo.

4 ESTUDIO DE MEDIDAS DE CALIDAD Y SU IMPACTO EN RECONOCIMIENTO DE LOCUTOR.

4.1 Marco experimental

En esta sección se explican las bases de datos, protocolos y sistemas utilizados en los experimentos de este PFC.

4.1.1 Bases de datos y protocolos

El organismo estadounidense NIST organiza planes de evaluación bianuales de carácter competitivo, para los cuales elabora bases de datos de voz y define una serie de protocolos experimentales. Parte de los experimentos de este trabajo están elaborados con estos protocolos y bases de datos [SRE 2006;SRE 2008].

La base de datos **NIST SRE 2006** está definida en detalle en [Campbell *et al.* 2004] . En este trabajo se utilizó la tarea *1conv4w-1conv4w,* que consiste en la utilización de 1 conversación para el entrenamiento y 1 conversación para testeo. Las locuciones tanto de testeo como de entrenamiento tienen una duración de 5 minutos con una duración media de 2,5 minutos tras eliminar los silencios. Los datos provienen de teléfonos fijos, inalámbricos y celulares.

La base de datos y protocolos **NIST SRE 2008** [NIST SRE 2008**]**también fueron elaborados para la evaluación de sistemas de reconocimiento de locutor. Está compuesta de dos tipos principales de habla: habla microfónica y habla telefónica, tanto para locuciones de testeo como para entrenamiento de modelos. Las condiciones de testeo y entrenamiento de modelos incluyen conversaciones grabadas sobre un canal telefónico, y sobre un canal microfónico en el escenario conocido como *interview* o de entrevista (en el cual existe el locutor principal y un entrevistador que formula preguntas), y adicionalmente locuciones de testeo grabadas sobre un canal microfónico. El protocolo establecido para la evaluación define las siguientes condiciones de entrenamiento: 10 segundos, 1 (short2), 3 y 8 conversaciones y *long conversation*, y las siguientes condiciones de testeo: 10 segundos, (1 short3) y *long conversation*. Cada conversación *short* tiene una duración media de 5 minutos, con una media de 2,5 minutos de habla (una vez eliminados los silencios). Las locuciones tipo interview contienen 3 minutos de habla registrada por micrófono, de los cuales la mayoría corresponde a habla del locutor y el resto a la voz del entrevistador.

Dentro del protocolo de evaluación se definen cuatro condiciones: *tlf-tlf, mic-mic, tlf-mic, y mic-tlf*. La primera de ellas es la utilizada en la sección 5.1 de experimentos con habla telefónica. Los resultados de las otras tres se muestran en la sección 5.2. Dentro de las locuciones microfónicas se utilizaron distintos tipos de micrófono con distintos grados de fidelidad, lo que implica una gran variabilidad de la calidad.

Para todas las condiciones de SRE 2008 se utilizó la tarea *short2-short3*, que consiste en una conversación para entrenamiento y otra para testeo.

Canal entrenamiento	Canal test	Número enfrentamientos	Target	Non-Target
Telefónico	Telefónico	37.050	3.832	33.218
	Microfónico	15.771	3.969	11.802
Microfónico	Telefónico	11.741	1.105	10.636
	Microfónico	34.046	11.525	22.521

Tabla 4.1*: número de enfrentamientos por tipo de canal y según Target y Non Target.*

Ahumada III [Ramos *et. al.*, 2008] es una base de datos forense que fue adquirida por el departamento de Procesamiento de Imagen y Acústica de la Guardia Civil. Incluye datos de locuciones de conversaciones reales que fueron registradas entre los años 1995 y 2004.

Las locuciones pertenecen a 61 locutores distintos. Existe una locución de entrenamiento por cada locutor, y tienen una duración de 2 minutos provenientes de una misma locución, entre los cuales no se encuentran silencios. Hay 5 locuciones de testeo para 30 de los locutores y 10 locuciones para 31 de ellos. Todas ellas con una duración de entre 7 y 25 segundos, con una duración media de 13 segundos.

Las grabaciones provienen tanto de teléfonos fijos como de terminales GSM y presentan una gran variabilidad en cuanto a entornos, estados emocionales, ruido ambiental, etc.

Esta base de datos presenta la peculiaridad de que en el formateo de la misma se suprimieron los silencios, quedando sólo las zonas de voz.

4.1.2 Sistemas

Para todos los experimentos de este trabajo se utilizaron los siguientes sistemas:

• **Sistema ATVS GMM**. Sistema basado en modelos GMM de 1024 mezclas y 19 parámetros MFCC, adaptados de un UBM entrenado con una gran cantidad de datos provenientes de evaluaciones NIST [NIST SRE] anteriores a 2006. Incluye compensación de canal con *feature warping* y *factor analysis.* Este sistema ha sido utilizado en [Ramos *et. al.*, 2008].

Para experimentos con bases de datos telefónicas se somete a estas a un filtrado de Wiener. De este modo los resultados mejoran notablemente que sin el filtrado.

• **Sistema SVM-GLDS.** Sistema SVM que hace uso del kernel GLDS para transformar los vectores de parámetros a espacios de dimensiones mayores. En [Mateos, 2007] se encuentra una implementación de este sistema.

• **Sistema SVM-SV.** Este sistema calcula un modelo GMM para cada locución involucrada en una comparación para después transformarlos en un supervector tal y como se explica en la sección 2.4.4. Sistema también implementado en [Mateos, 2007]

4.2. Metodología

4.2.1 Estructura

En esta sección se presenta la metodología utilizada en el desarrollo de los experimentos de este trabajo.

En primer lugar, se distinguirán cuatro grupos principales de experimentos, en función de los tipos de datos con los que se ha realizado el estudio:

1. **Telefónicos**: tanto las locuciones de entrenamiento como de testeo proceden de grabaciones telefónicas.
2. **Microfónicos**: ambas locuciones fueron registradas con micrófonos, sin haber sido sometidas a transmisión por un canal de telecomunicación.
3. **Microfónico-telefónicos (o cruces)**: una de las locuciones que intervienen en cada prueba es de tipo microfónico, y la otra de tipo telefónico.
4. **Bases de datos forenses**: todas las locuciones provienen de grabaciones telefónicas de conversaciones reales, de diferentes tipos de terminales (fijos, inalámbricos, GSM…).

En los cuatro casos hay diferentes factores que pueden variar la calidad de la señal de voz, dependiendo de diversos factores ya citados en el apartado 3.1. En el caso de las señales telefónicas, depende primordialmente de las distorsiones introducidas por el canal de transmisión. Para las microfónicas no existe este problema, sin embargo puede verse afectado por otros factores como las condiciones de registro (distancia al locutor, actitud del locutor, tipo de micrófono, condiciones ambientales…). En bases de datos forenses, a estos factores ya comentados hay que añadir la falta de control sobre las condiciones de registro (entorno, distintos dispositivos de registro, estado emocional del locutor), por lo que la calidad en este tipo de bases de datos suele tener mayor variabilidad.

Dentro de cada uno de estos tres grupos, la metodología seguida ha sido la misma, y se resume en los siguientes puntos:

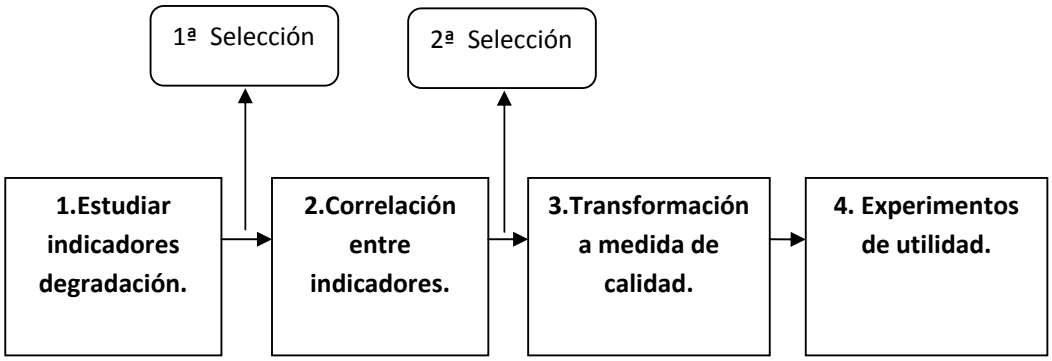


Fig 4.1: *esquema de la metodología experimental utilizada.*

1. **Estudiar indicadores de degradación (en adelante IDs)**: se extraerán de las locuciones una serie de medidas (algunas encontradas en el estado del arte, otras propuestas en este trabajo), y se probará su eficacia para determinar la degradación de las locuciones, utilizando como criterio el rendimiento de los sistemas**.**

2. **Correlación entre indicadores**: consiste en estudiar en qué grado es complementaria la información que aportan los IDs. Esta información permitirá combinarlos de manera óptima cuando se hagan trabajos de compensación de calidad (no se hará en este trabajo), además de servir para descartar algunos indicadores para poder centrarnos en aquellos más prometedores.
3. **Transformación a medida de calidad**: a partir del estudio de cada indicador, se llevará a cabo un mapeo que defina una medida de calidad tal y como recomienda el estándar de NIST [Alonso *et. al.,* 2008,Grother *et. al.,* 2008].
4. **Experimentos de utilidad:** permitirán observar en qué grado pueden ayudar las medidas a predecir el rendimiento de un sistema, y por lo tanto su utilidad para compensar la calidad.

A continuación se pasa a describir con la metodología con mayor detalle.

4.2.2 Estudio de los indicadores de degradación.

En este apartado se explica el procedimiento seguido para observar la relación entre la magnitud de los distintos IDs y el rendimiento de los SRL. Dicho experimento no se ha encontrado en ningún otro estudio en la materia, pero se propone en este trabajo como método para extraer la mayor información posible sobre los indicadores de degradación. Por lo tanto este método constituye una contribución importante del presente PFC.

El método propuesto básicamente consiste en medir el rendimiento de un sistema, para varios subconjuntos de locuciones con distintos valores de ID. En los siguientes puntos se explican los pasos seguidos para llevar a cabo el experimento:

1. Se calcula el valor de ID para cada locución de la base de datos.
2. Se calculan las puntuaciones de similitud para el sistema y base de datos escogido.
3. Para cada comparación "*i*", se calcula la magnitud media del ID "μ_i" como la media aritmética de los dos IDs que intervienen en la comparación.
4. Los scores son ordenados conforme a su μ_i de menor a mayor.
5. Se selecciona un conjunto del 20% de scores con ID más bajo, que llamaremos conjunto *k*, y se obtiene el EER correspondiente a este conjunto y el ID medio del conjunto.
6. Este último paso se repite 100 veces, para los conjuntos de scores K=1…100, cada vez con mayor ID medio, que irá aumentando siempre en la misma proporción. El último conjunto contendrá el 20% de los scores con mayor ID medio.
7. Finalmente obtenemos un vector de 100 valores de EER para los 100 subconjuntos de scores solapados. Dicho EER se representará gráficamente frente al valor ID medio de los scores. Dichas curvas las denominaremos curvas "Magnitud vs Rendimiento".

El resultado es una gráfica como la que se muestra a continuación, en la cual se observa la relación entre la magnitud del ID y el rendimiento del sistema (en este caso EER en %). Corresponde al ID Skewness Local, y en este caso se muestran tres curvas correspondientes a los sistemas GMM, SV y GLDS:

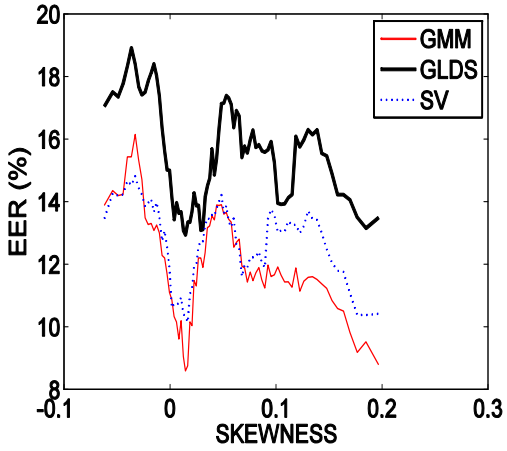


Fig. 4.2: ejemplo de curva "Magnitud vs Rendimiento" para el ID Skewness Local.

Para analizar estas gráficas, se tienen en cuenta una serie de criterios:

• **Impacto:** una medida será representativa de la degradación de las locuciones si existe una relación entre la magnitud de dicha medida y el rendimiento del sistema. El impacto de dicho ID se define como la diferencia máxima entre los valores de rendimiento para todo el rango de variación del ID dividido entre el máximo valor de EER. Por ejemplo, en la gráfica que se ha mostrado, para la curva de color rojo (en este caso correspondiente al sistema GMM) el impacto sería:

$$Impacto = \frac{EER_{MAX} - EER_{MIN}}{EER_{MAX}} \cong \frac{16 - 8}{16} = 50\%$$

Para cada sección de experimentos se resumirá en una tabla los valores de impacto de todos los IDs. Es importante resaltar que este valor por sí mismo da una idea de utilidad sólo parcialmente, ya que es posible que la curva sufra grandes oscilaciones sin mantener una tendencia coherente, y en tal caso un descenso grande del EER no sería indicativo de un alto impacto. Por lo tanto la información de estas tablas siempre será complementada analizando la tendencia de las curvas EER vs ID.

• **Coherencia entre sistemas y entre bases de datos**. Pese a que se espera que la curva de EER vs magnitud de un ID difiera para distintos sistemas y bases de datos, el hecho de que guarden cierta similitud representa un apoyo a la validez del ID propuesto.

Una vez dispongamos de estos resultados, se hará una **primera selección** de los IDs con los que se continuará el estudio. Se descartarán aquellos en los que se observe una menor utilidad como indicadores de degradación.

4.2.3 Estudio de correlación entre indicadores.

El estudio de la correlación de los indicadores consiste en conocer cuantitativamente y cualitativamente el grado en que se complementa la información aportada por dichos indicadores. Por ejemplo, si tuviéramos que combinar dos IDs para determinar la calidad de una locución, sería lógico pensar que el ID SNR-Wiener y el SNR proporcionaran información muy parecida y sería preferible combinar sólo uno de ellos con otro indicador distinto, como el P563 que proporciona información más complementaria.

Los resultados que se obtengan en este paso nos servirán para descartar parte de los indicadores y centrarnos en unos pocos para los experimentos de utilidad.

Los experimentos de correlación se representan en gráficas "Scatter-Plots", una para cada combinación de dos indicadores. De este modo, a partir de la forma de la nube de puntos se puede interpretar como están relacionadas ambas medidas. Además se añadirá su coeficiente de correlación lineal en la parte superior de la gráfica, que si es cercano a uno en valor absoluto indica muy alta correlación, y si es cercano a cero indica muy baja correlación. A continuación se muestra un ejemplo para los IDs UBML y SNR.

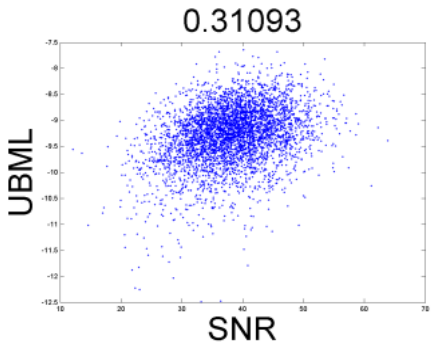


Fig. 4.3: *Scatter plot de correlación de ejemplo.*

4.2.4 Transformación a medida de calidad

En este paso se aplicará una función de transformación a cada ID, de modo que quede una medida de calidad, definida como recomienda el organismo de estandarización NIST [Grother *et. al.,* 2008]. De este modo, para cada valor de ID "x", se definirá una función de transformación $Q(x)$ que tomará valores entre 0 y 1, correspondientes a la peor y mejor calidad posible respectivamente.

En algunos casos la función de mapeo será determinada por la información disponible a priori sobre el ID, como en el P563, para el cual siempre se generan valores entre 1 y 5 (la peor y mejor calidad respectivamente). En otros casos, no disponemos información a priori por lo que nos basaremos en el estudio de los IDs (curvas Magnitud vs Rendimiento) para determinar la función.

Es importante resaltar que dicho mapeo no puede considerarse absoluto, ya que ha sido diseñado basándose en las curvas "Magnitud vs Rendimiento", obtenidas a partir de dos bases de datos (NIST SRE 2006 y SRE 2008), ambas con niveles de calidad similar.

A continuación se muestra un ejemplo de mapeo a medida de calidad, correspondiente al ID KLPC para la base de datos SRE 2006 y los tres sistemas estudiados:

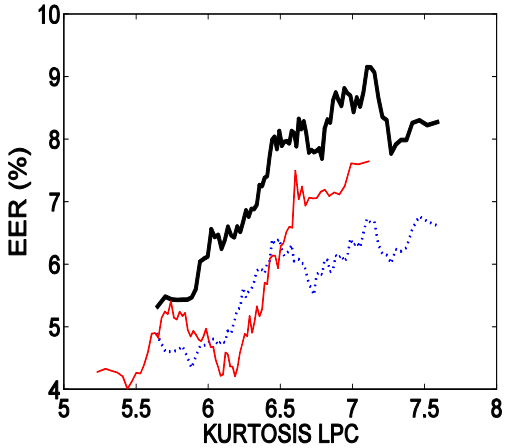


Fig.4.4: *ejemplo de curva Magnitud vs Rendimiento para el ID Kurtosis LPC*

Como se observa hay una clara tendencia a aumentar el EER según aumenta la KLPC. Para definir la función se deben establecer unos límites y el tipo de aproximación (lineal, no lineal, creciente o decreciente). Los límites serán establecidos a partir de la distribución de los IDs para las locuciones, mientras que el tipo de aproximación se tomará de la propia curva. El resultado en este caso sería:

Indicador	Rango	Aproximación	Función de mapeo
Kurtosis LPC	[3,11]	Lineal decreciente	$Q_{KLPC}(x) = 1 - \left(\frac{x-3}{8}\right)$

Tabla 4.2*: ejemplo de función de mapeo a medida de calidad para el ID Kurtosis LPC*

4.2.5 Experimentos de utilidad

Los experimentos que se presentan en esta sección tienen como objetivo mostrar la efectividad de las mediadas de calidad para predecir el rendimiento de un sistema. Para ello se muestran tres tipos de gráficas:

• Curvas "Score vs Calidad"

Son diagramas de dispersión (*Scatter-Plots*) de la puntuación de similitud frente a la medida de calidad, desglosadas por puntuaciones target y puntuaciones non-target. En el eje horizontal se representa la medida de calidad de la puntuación, calculada como la media geométrica de las calidades de entrenamiento y de test (sección 3.2.3) y en el eje vertical se representa la puntuación de similitud. De este modo, se pretende observar como varía la capacidad discriminativa del sistema para distintos valores de calidad. Si la medida de calidad es útil, se espera que las nubes de puntos Target y NonTarget estén más separadas para

mejores valores de calidad. En otras palabras, el sistema debería discriminar mejor los scores cuando estos poseen una calidad más alta. Para poder observar mejor la tendencia de las nubes, se añaden rectas de regresión a las mismas.

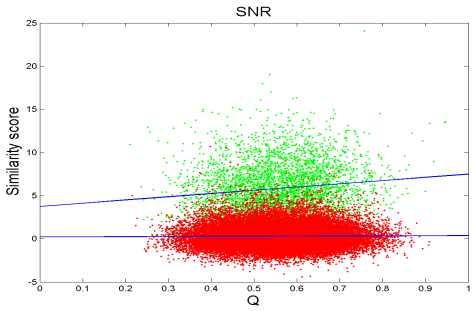


Fig. 4.5: *ejemplo de scatter-plot "Score vs Calidad" para el ID SNR*

• Curvas DET

Estas curvas son muy populares para representar el rendimiento de un sistema [Martin *et. al.,* 1997]. Básicamente consiste en representar la relación entre los dos tipos de errores que puede cometer un sistema (falsa aceptación y falso rechazo). En este caso las utilizaremos para representar el rendimiento de un sistema cuando utilizamos distintos subconjuntos de comparaciones de un experimento con distintos valores medios de calidad.

Más concretamente, el método utilizado consiste en representar la curva DET para dos conjuntos de comparaciones: por un lado la población entera para un experimento dado, y por otro lado el 75% de las comparaciones con mejor calidad media. La calidad media de la comparación será calculada de la misma manera que en apartado anterior (mediante la media geométrica de ambas locuciones). De esta manera se espera que la curva DET del segundo conjunto se acerque al origen, disminuyendo el EER.

El resultado sería una gráfica como la que se muestra a continuación, en la cual se observan seis curvas, correspondientes a las dos curvas mencionadas para tres sistemas: SVM-GLDS, SVM-SV y GMM.

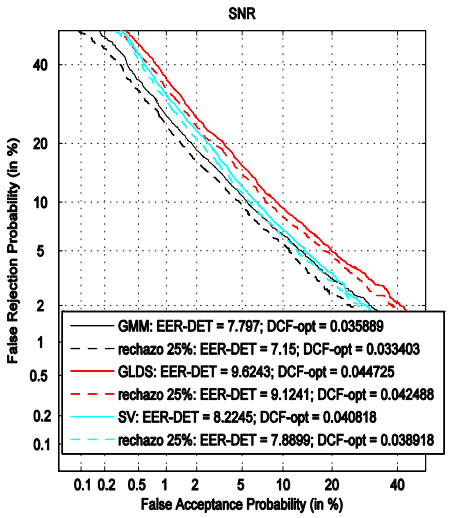


Fig 4.6*: ejemplo de curva DET con rechazo del 25% de scores con peor calidad*

La medida de calidad será tanto más útil para predecir el rendimiento en discriminación cuanto más mejore la curva DET al rechazar el 25% de comparaciones con peor calidad.

• Curvas "Error vs Exclusión"

En este tipo de curvas se pretende observar las variaciones del EER según se excluyen fracciones cada vez mayores de un conjunto de scores. De este modo, no sólo se cuantifica la utilidad de la medida, como se hacía con las curvas DET, sino que también conocemos como es la progresión de la mejora del rendimiento. El NIST recomienda estas curvas para mostrar la utilidad de medidas de calidad [Alonso *et. al.* 2008].

La tendencia de estas curvas depende de varios factores, que tendrán que ser tenidos en cuenta a la hora de extraer las conclusiones:

• **Distribución de la medida de calidad**: dependiendo de ésta, la calidad media del conjunto de scores variará más o menos bruscamente tras la exclusión de una fracción de scores.

• **Relación entre la medida y el rendimiento del sistema**. Dicha relación puede variar dependiendo del sistema y la medida de calidad. La mayoría de los casos se aproxima a una función lineal.

• **Mapeo**. Un mapeo erróneo puede dar lugar a tendencias que no reflejen la utilidad real de una medida.

A continuación se muestra un ejemplo de este tipo de curvas.

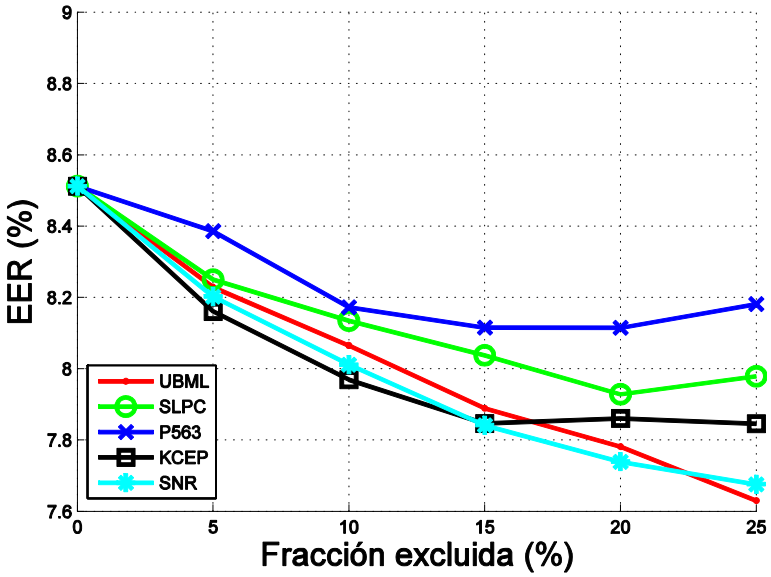


Fig 4.7: *ejemplo de curvas "Error vs Exclusión".*

5 RESULTADOS

5.1 Medidas de calidad en bases de datos telefónicas.

En este apartado se muestran los resultados obtenidos con las bases de datos telefónicas SRE 2006 1conv4w-1conv4w (habla telefónica) y las locuciones telefónicas de SRE 2008. Se ha experimentado con los sistemas SVM Super Vector, SVM GLDS y GMM (Sección 4.1).

5.1.1 Estudio de indicadores de degradación.

Con los resultados aquí presentados se pretende conocer en detalle la relación entre los distintos indicadores con el rendimiento de cada uno de los sistemas.

A continuación se muestran las gráficas EER vs Magnitud para los indicadores de degradación estudiados (explicados en la sección 3.3). En cada gráfica se han introducido tres curvas (una por sistema) y las gráficas de una misma fila corresponden a un mismo indicador de degradación, así se facilita la comparación entre sistemas y bases de datos, acorde con los criterios de coherencia establecidos en la sección 4.2.

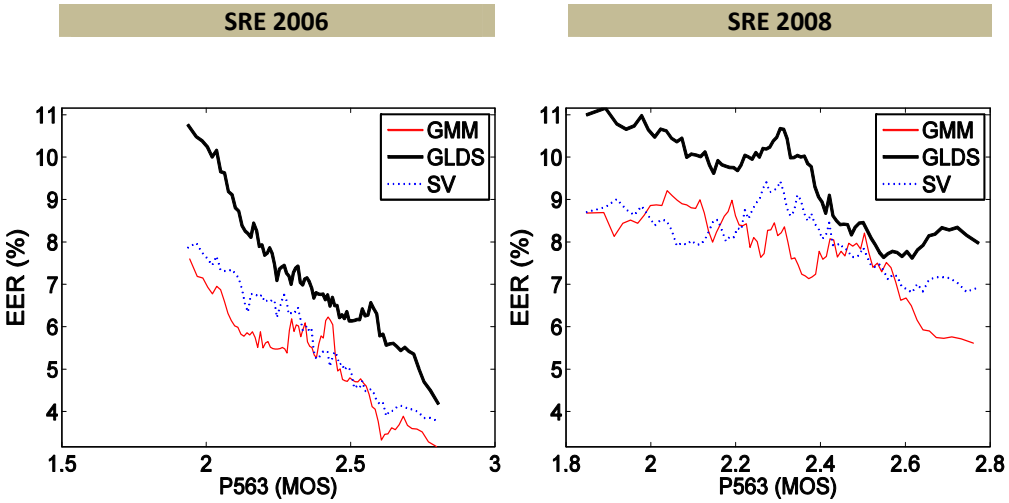


Fig. 5.1.a. Gráficas EER (%) vs Magnitud del indicador de degradación para los IDs indicados por el rótulo bajo las figuras

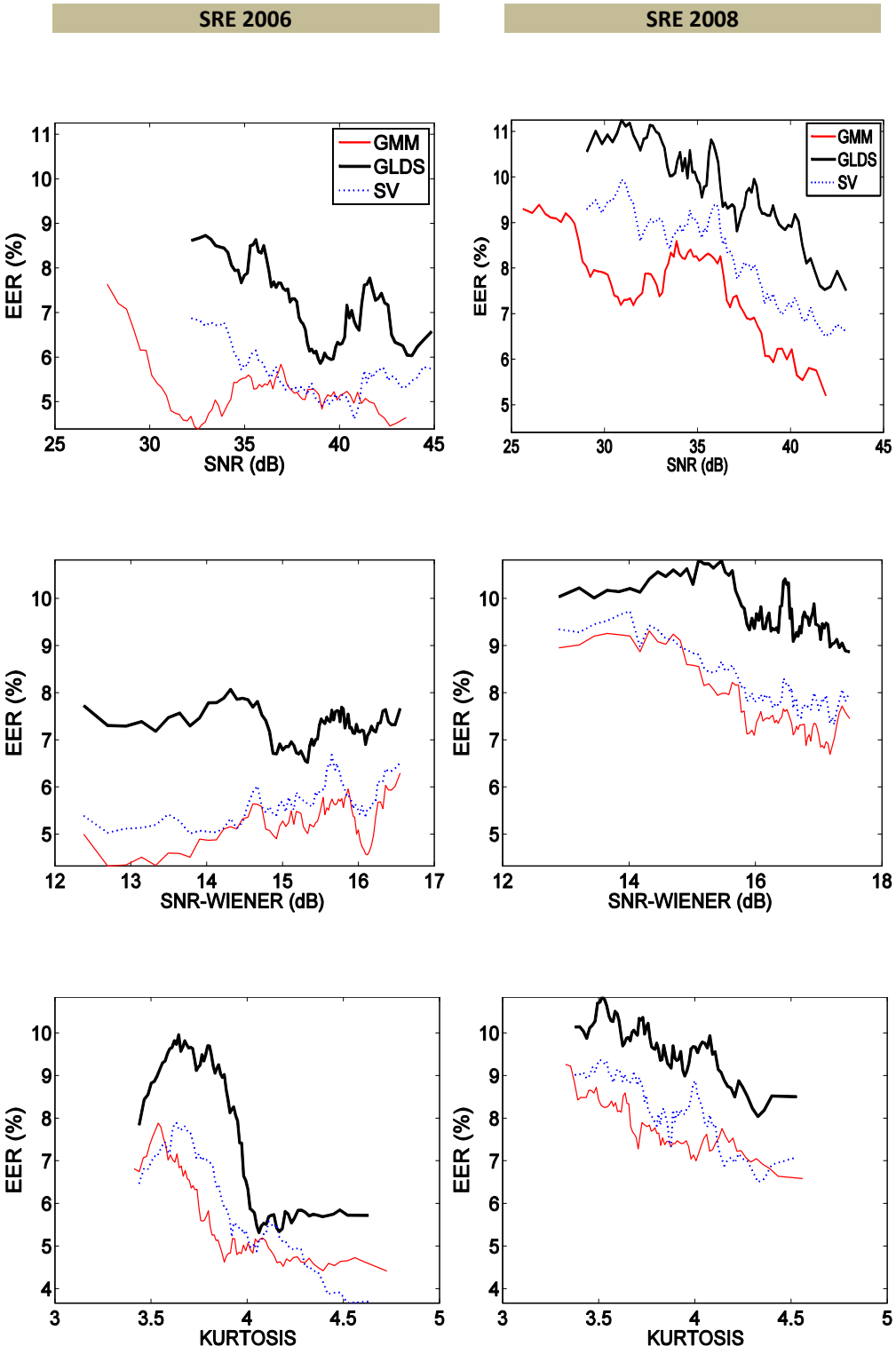


Fig. 5.1.b. *Gráficas EER (%) vs Magnitud del indicador de degradación para los IDs indicados por el rótulo bajo las figuras*

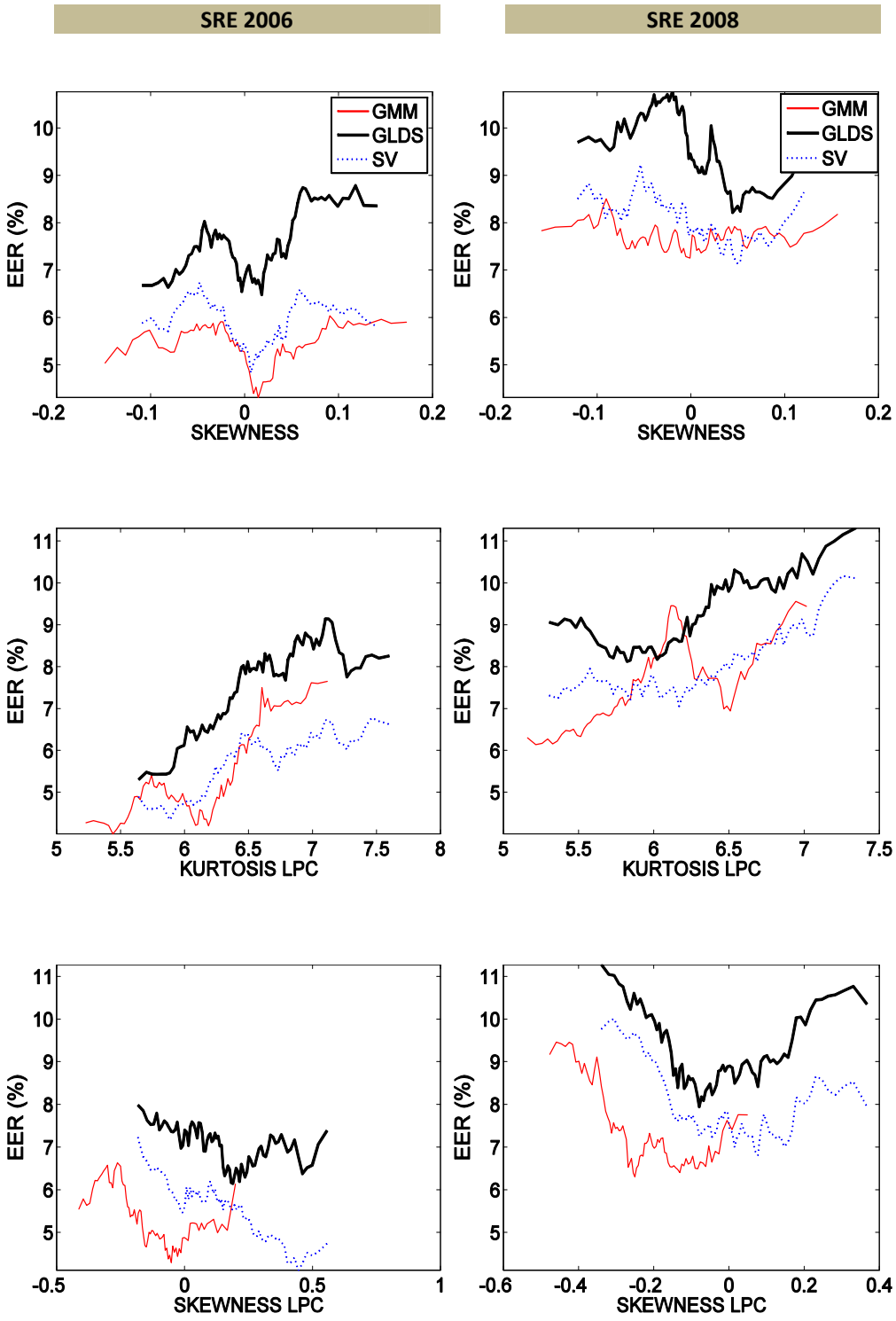


Fig. 5.1.c: *Gráficas EER (%) vs Magnitud del indicador de degradación para los IDs indicados por el rótulo bajo las figuras.*

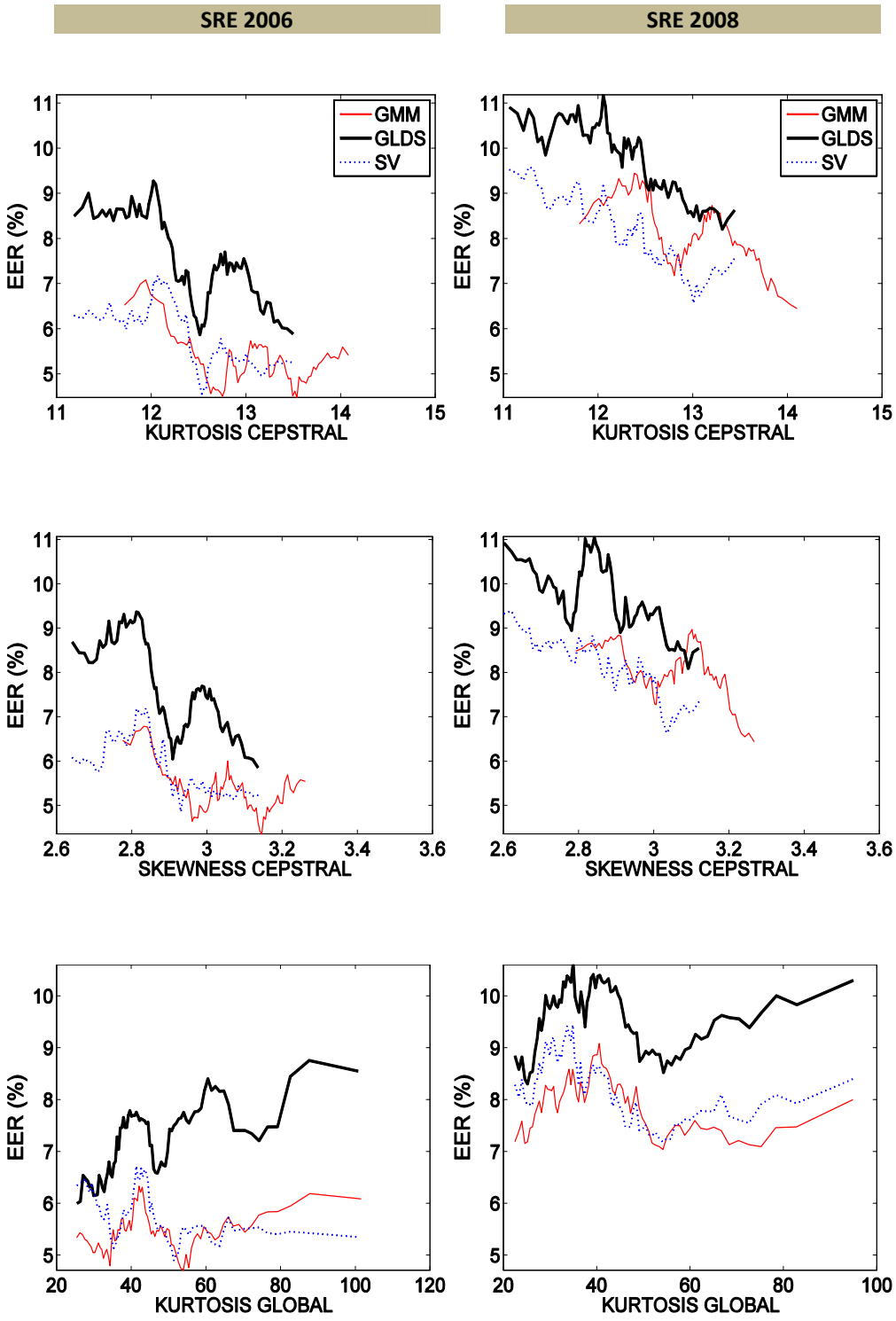


Fig. 5.1.d: *Gráficas EER (%) vs Magnitud del indicador de degradación para los IDs indicados por el rótulo bajo las figuras.*

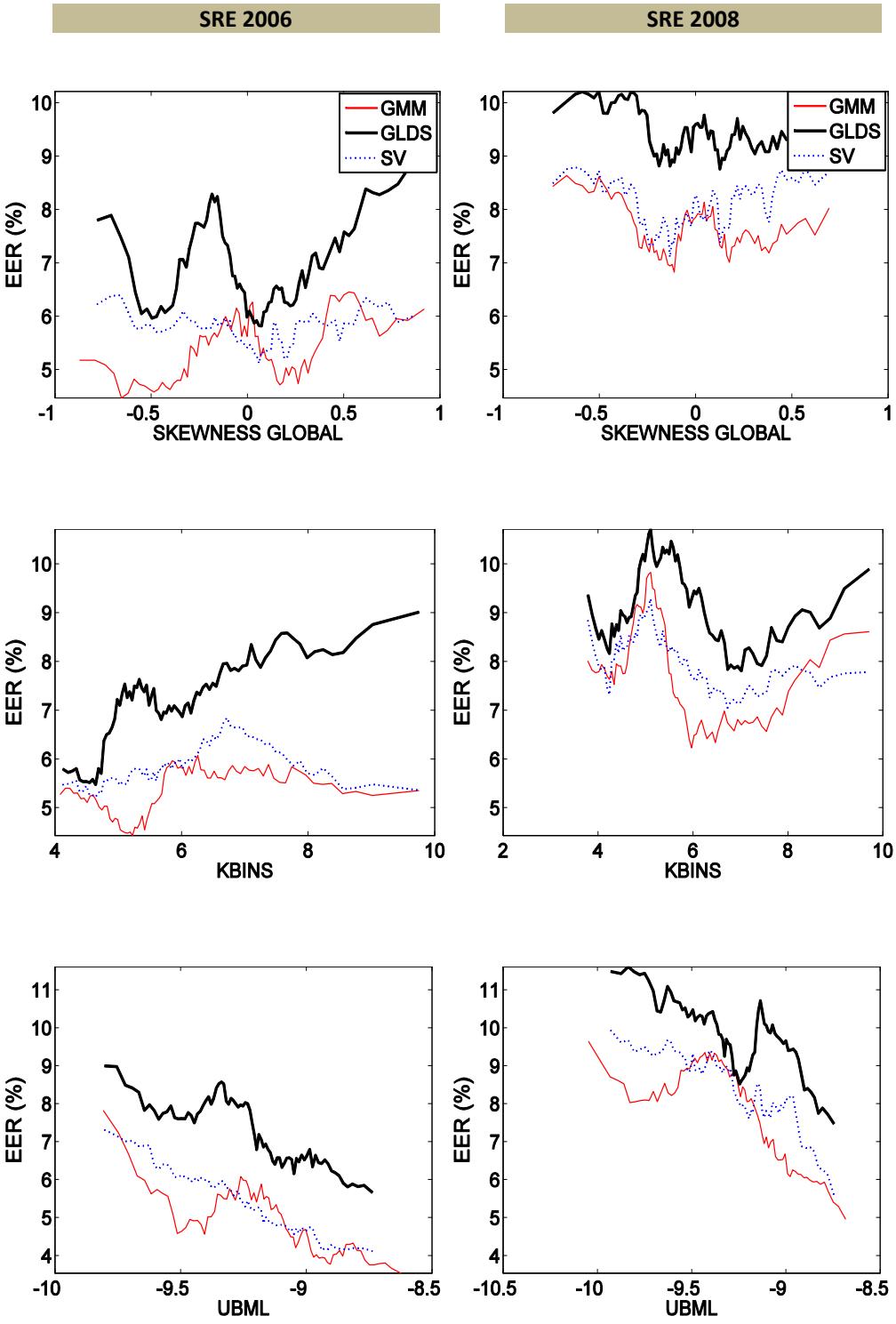


Fig. 5.1.e: *Gráficas EER (%) vs Magnitud del indicador de degradación para los IDs indicados por el rótulo bajo las figuras.*

Las conclusiones extraídas de las gráficas se resumen en los siguientes puntos:

• **Relación entre el rendimiento y la magnitud del indicador (impacto)**. Para gran parte de los indicadores se observa una clara relación entre el EER y la magnitud de los mismos, exceptuando algunos en los que la tendencia de las curvas no es tan clara (skewness,SNR-Wiener,kurtosis global,skewnes global,Kbins). En la siguiente tabla se muestran los porcentajes de mejora de cada indicador de degradación en %. Es muy importante recalcar que esta mejora sólo será significativa si en la tendencia de las gráficas se aprecia una clara relación entre el ID y el rendimiento del sistema. De este modo, no se identificarán como alto impacto aquellos valores altos de mejora cuyas curvas no tengan una tendencia clara.

Sistema / Indicador	SRE 2006			SRE 2008			Media
	GLDS	SV	GMM	GLDS	SV	GMM	
P563	61%	52%	58%	30%	25%	39%	**44%**
SNR	32%	17%	38%	31%	30%	41%	**32%**
SNR-WIENER	19%	25%	31%	18%	24%	28%	**24%**
KURTOSIS	32%	39%	30%	26%	24%	26%	**30%**
SKEWNESS	28%	26%	28%	15%	23%	22%	**24%**
KURTOSIS LPC	47%	31%	46%	31%	27%	34%	**36%**
SKEWNESS LPC	23%	43%	33%	26%	29%	30%	**31%**
KURTOSIS CEPSTRAL	37%	37%	37%	27%	31%	32%	**34%**
SKEWNESS CEPSTRAL	36%	36%	37%	26%	31%	32%	**33%**
KURTOSIS GLOBAL	31%	27%	26%	27%	24%	37%	**29%**
SKEWNESS GLOBAL	20%	15%	16%	14%	19%	20%	**17%**
KBINS	38%	23%	27%	27%	24%	36%	**29%**
UBML	37%	44%	55%	36%	44%	50%	**44%**
	34%	**32%**	**36%**	**26%**	**27%**	**33%**	

Tabla 5.1: *mejora del EER (en %) para cada indicador de degradación*

Por lo general se observa un mayor impacto para la base de datos SRE 2006. En cuanto a los indicadores, se observa que dos de ellos destacan por tener un impacto notablemente mayor al resto: P563 y UBML, que es la que se ha propuesto en este PFC.

• **Coherencia entre sistemas y entre bases de datos**: por lo general se observa un comportamiento coherente, tanto entre sistemas como entre las dos bases de datos, lo cual refuerza la validez de los resultados obtenidos. Una observación que cabe resaltar, es el desplazamiento de la curva del sistema GMM con respecto a los otros dos sistemas. La base de datos con la que se generaron los resultados del sistema GMM contiene las mismas locuciones pero filtradas, lo cual puede explicar tal desplazamiento.

Para tener una visión global de cada una de las medidas, en la siguiente tabla se resumen las características de cada uno de los indicadores. Se ha estructurado en tres columnas: impacto, coherencia entre sistemas y coherencia entre bases de datos. Para cada una de estas columnas

se ha calificado el indicador con tres posibles valores (+, o , -), que indican una calificación alta, media o baja.

Indicador	Impacto	Coherencia sistemas	Coherencia bases datos
P563	**+**	**+**	**+**
SNR	**+**	**+**	**+**
SNR-WIENER	o	-	-
KURTOSIS LOCAL	+	+	o
SKEWNESS LOCAL	-	-	-
KURTOSIS LPC	o	o	+
SKEWNESS LPC	o	o	+
KURTOSIS CEPSTRAL	o	o	+
SKEWNESS CEPSTRAL	o	o	+
KURTOSIS GLOBAL	-	-	-
SKEWNESS GLOBAL	-	-	-
KBINS	-	-	-
UBML	**+**	**+**	**+**

Tabla 5.2*: calificaciones de los indicadores de degradación.*

Por lo general se observa que un mayor impacto implica una mayor coherencia entre sistemas y entre bases de datos. Observamos que hay tres indicadores que destacan por sus características: P563, SNR y UBML..

5.1.2 Experimentos de correlación

El estudio de los IDs ha permitido identificar aquellos que proporcionaban mayor información sobre la calidad del audio, para poder descartar aquellos que aporten menos, y centrarnos en los más prometedores. Con el siguiente experimento se pretende conocer cuán correlados están dichos indicadores, lo cual servirá para hacer una segunda selección y descartar aquellos que no aporten información complementaria. Estos han sido los indicadores seleccionados:

- P.563
- SNR
- Kurtosis LPC (“KLPC”)
- Kurtosis local en el dominio temporal (“KURTOSIS LOCAL”)
- Kurtosis en el dominio Cepstral (“KCEP”)
- Skewness en el dominio Cepstral (“SCEP”)
- Verosimilitud con UBM (“UBML”)

En las siguientes figuras podemos observar los diagramas de dispersión (*Scatter-Plots*) de las medidas mencionadas, junto con sus coeficientes de correlación lineal, para las bases de datos SRE 2006 y SRE 2008.

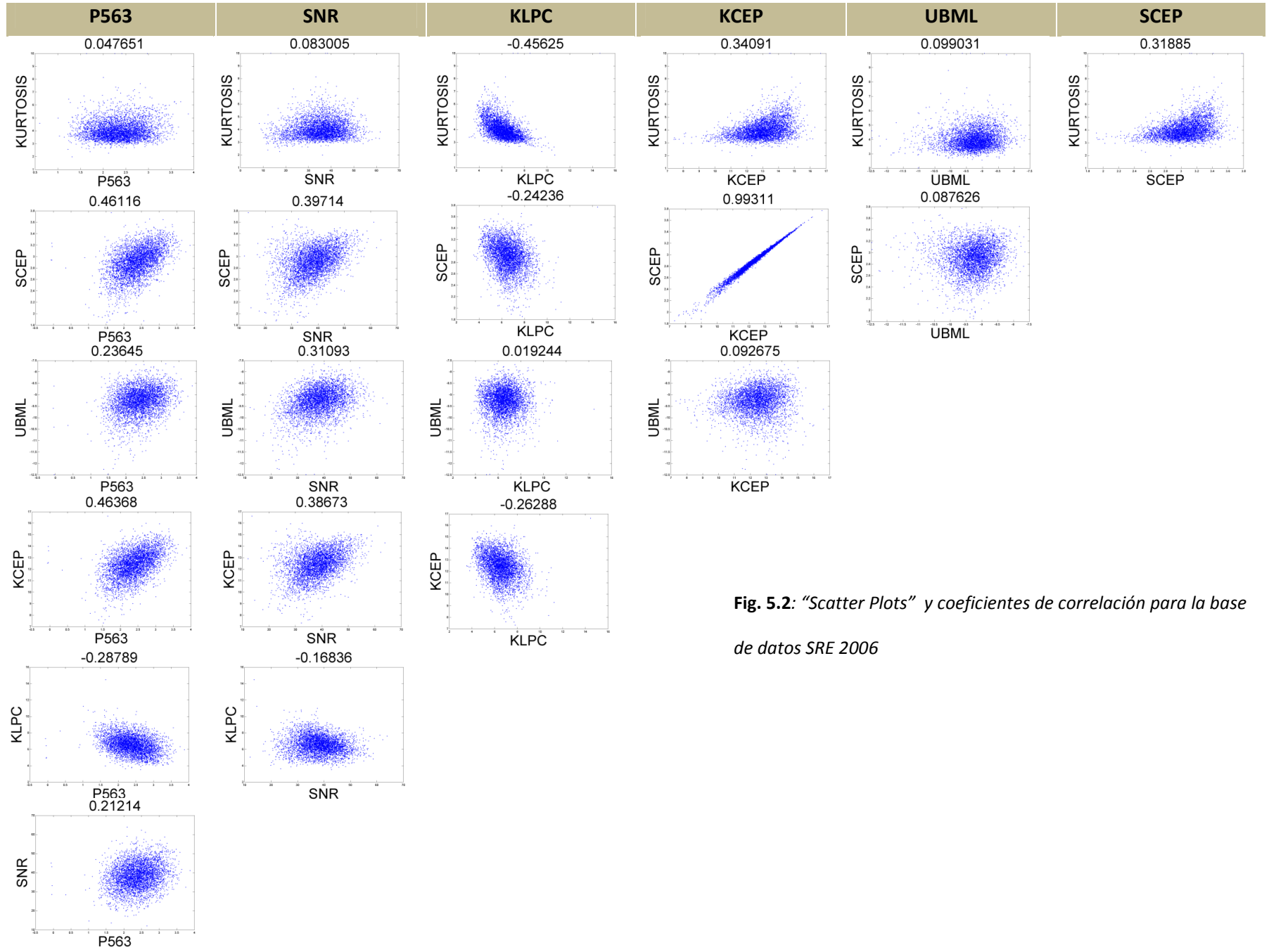


Fig. 5.2: *"Scatter Plots" y coeficientes de correlación para la base de datos SRE 2006*

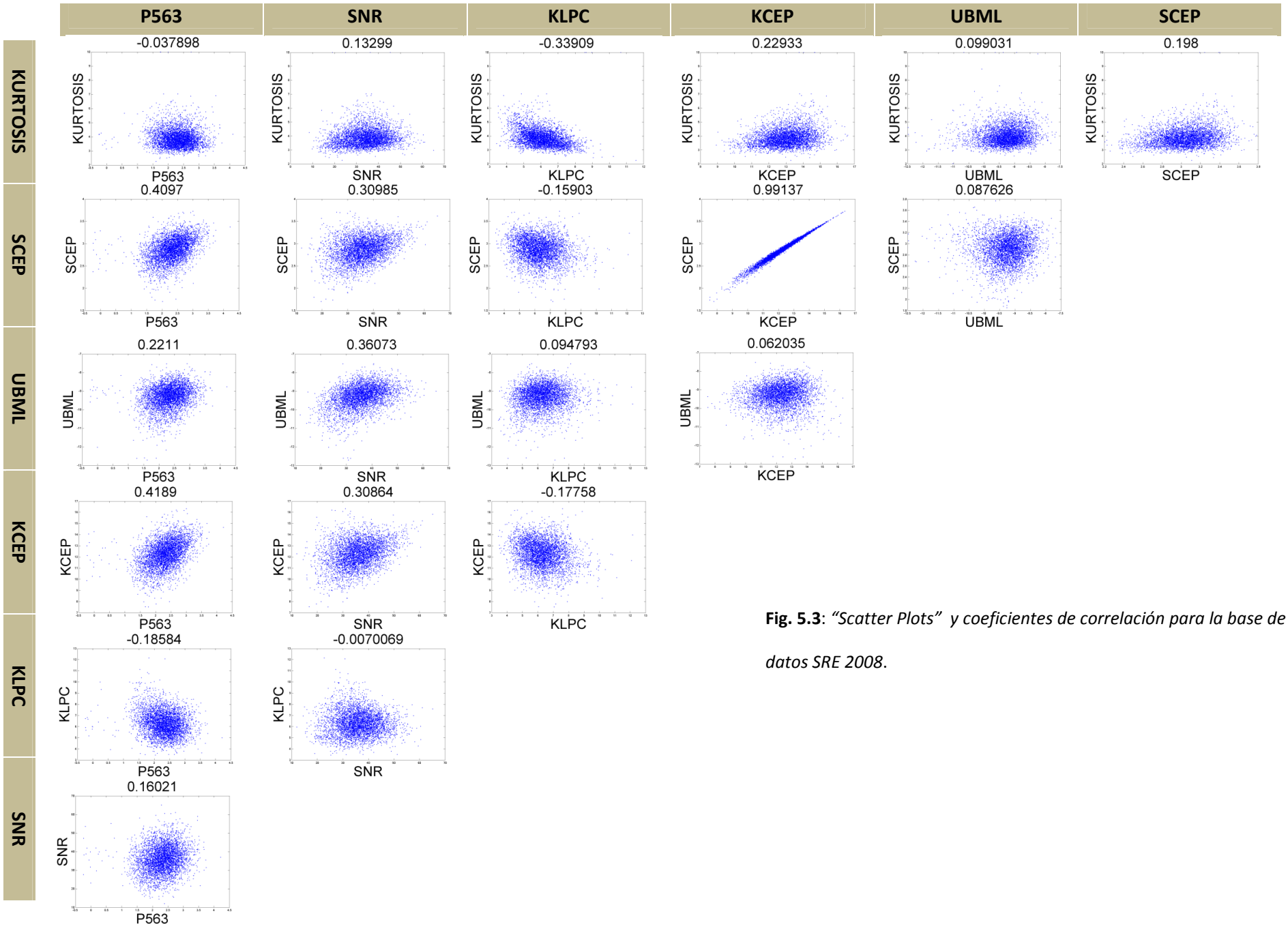


Fig. 5.3: *"Scatter Plots" y coeficientes de correlación para la base de datos SRE 2008.*

Se observa que los coeficientes de correlación son coherentes entre las dos bases de datos.

La kurtosis muestra una correlación muy baja con respecto a los indicadores P563, SNR y UBML. Por lo tanto proporciona información muy complementaria a dichos IDs.

El coeficiente de correlación entre los IDs P563 y SNR es bajo para ambas bases de datos. Probablemente debido al bajo nivel de ruido en estas, ya que el algoritmo P563 toma algunos indicadores clave para determinar el tipo de degradación principal a la que está sometido el audio (entre ellas la SNR). Al estar poco correlados, se puede entender que el algoritmo P563 descarta la SNR como factor determinante de la calidad, y le otorga muy poco peso en el cálculo de la misma. Esta hipótesis cobró mayor sentido al calcular la media de la distribución de la SNR, que en ambas bases de datos es superior a los 30 dB (Anexo A).

Se observa una correlación lineal muy alta entre las medidas estadísticas en el dominio Cepstral: KCEP y SCEP. Por lo tanto la información que proporcionan es altamente redundante, por lo que se podría descartar una combinación de ambas para obtener mayor información sobre la calidad del audio. Por este motivo, en los experimentos de utilidad que serán vistos en la siguiente sección, se descartará el indicador "Skewness Cepstral".

La medida P563 muestra una correlación moderada con el resto de medidas, excepto con la kurtosis. Como se citó en el apartado 3.3, el algoritmo P563 toma como parámetros la SNR, KLPC, KCEP y SCEP, entre otros, para determinar la calidad del audio, por lo que estos valores parecen lógicos. La kurtosis sin embargo, muestra una correlación muy baja con la medida P563, lo cual podría indicar que no está relacionada con la calidad subjetiva del audio (la percibida por el oído humano) pero sí con la calidad de cara a un sistema reconocedor de locutor.

Cabe destacar los valores que se observan para la nueva medida propuesta, UBML. Se observa que mantiene una correlación notable (entre 0,3 y 0,4) con la medida SNR. Esto implica que la similitud de una locución con el modelo universal utilizado en la prueba, es sensible al nivel de ruido del audio. Si tenemos en cuenta el bajo nivel de ruido de la base de datos, el valor de correlación observado parece ser indicativo de una alta sensibilidad del UBML frente al nivel de ruido, ya que el UBM con el que fue calculada se entrenó con un bajo nivel de ruido.

5.1.3 Selección de medidas de calidad y mapeo

Con la información obtenida en el apartado anterior podemos hacer una segunda selección de los IDs a estudiar. A continuación se comentan los ID excluidos junto con su justificación:

- Skewness cepstral: al estar altamente correlado con la Kurtosis Cepstral, podemos considerar que la información que aporta es prácticamente la misma.
- Kurtosis: aunque se ha demostrado su utilidad como indicador de degradación, se ha observado que mantiene una correlación notable con el resto de medidas estadísticas, y su impacto no es tan alto como en el resto de indicadores seleccionados.

Los indicadores seleccionados, que por tanto serán transformados a medida de calidad, son los siguientes:

- KURTOSIS CEPSTRAL.

- UBML
- P.563
- SNR
- KURTOSIS LPC

El primer paso para estudiarlas como medida de calidad, es aplicarles una función de mapeo, para que estén limitadas al rango [0,1] tal y como se describe en la metodología de medidas de calidad propuesta[Gª Romero *et.al.* 2005, Grother *et. al.* 2008]. Las funciones fueron diseñadas en base a la información previa que disponíamos de cada indicador (en la mayoría de los casos nula) y a la forma de la curva Rendimiento vs Magnitud extraída. A continuación se muestran las funciones de mapeo que fueron asignadas a cada indicador.

Indicador	Rango	Aproximación	Función de mapeo
P563	(1,5)	Lineal creciente	$Q_{P563}(x) = \frac{(x-1)}{4}$
SNR	[0,60]	Lineal creciente	$Q_{SNR}(x) = \frac{x}{60}$
Kurtosis LPC	[3,11]	Lineal decreciente	$Q_{KLPC}(x) = 1 - \left(\frac{x-3}{8}\right)$
UBML	[-13,-5]	Lineal creciente	$Q_{UBML}(x) = \frac{(x+13)}{8}$
Kurtosis cepstral	[10,6]	Lineal creciente	$Q_{KCEP}(x) = \frac{(x-10)}{6}$

Tabla 5.3: *funciones de mapeo de indicador de degradación a medida de calidad, para los IDs P563, SNR, KLPC, UBML y KCEP.*

5.1.4 Experimentos de utilidad

Como se cita en la sección 4, los resultados de los experimentos de utilidad se muestran en dos tipos de gráficas: gráficas Score vs Calidad, que permiten observar cómo se separan los scores *Target* y *Non-Target* según incrementa la calidad, y las gráficas DET, que permiten observar la utilidad de la calidad como predictora del rendimiento del sistema.

• Gráficas Score vs Calidad

En las siguientes gráficas se representan el score de similitud (eje vertical) del sistema GMM (para las bases de datos SRE 2006 y SRE 2008), y el valor de la medida de calidad (Q) de la comparación o score (eje horizontal), calculado como la media geométrica de los dos valores de calidad que intervienen en cada comparación. La nube superior de puntos (de color verde

claro) representa comparaciones de usuarios, mientras que la nube inferior (color rojo) representa las comparaciones de impostores. Se han añadido líneas de regresión lineal a las nubes para poder observar mejor la tendencia de los scores target y non-target.

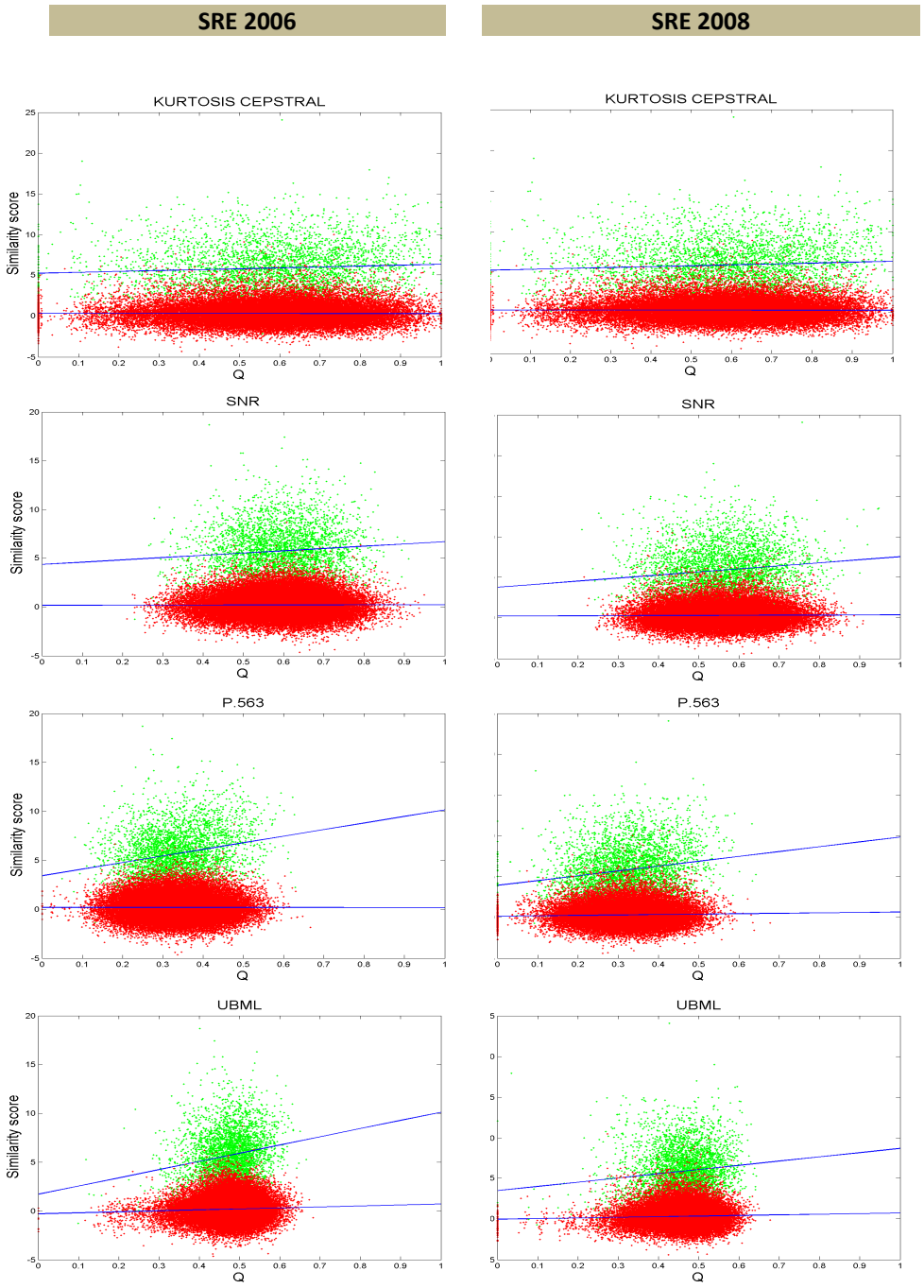


Fig. 5.4.a: *Gráficas Score vs Q, para las bases de datos SRE 2006 y SRE 2008.*

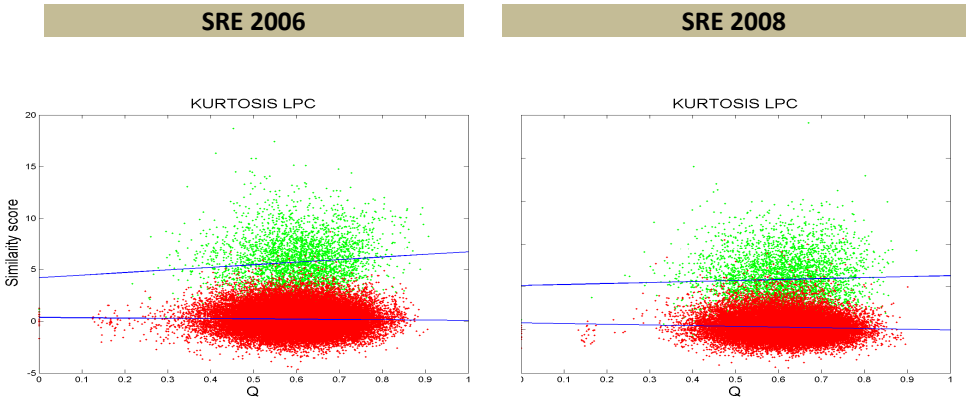


Fig. 5.4.b*: Gráficas Score vs Q, para las bases de datos SRE 2006 y SRE 2008.*

En todas las gráficas se observa que las nubes de puntos tienden a alejarse según incrementa el valor de la medida de calidad. Para cada una de las medidas el comportamiento es muy similar en ambas bases de datos, como ya se ha visto en el estudio de los indicadores de degradación, aunque en la base de datos SRE 2006 ese efecto es mayor.

Por lo general la nube que más se separa es la de comparaciones t*arget*, mientras que en la n*on-target* apenas se aprecia la separación. En otras palabras, en comparaciones *target* el sistema tiende a otorgar puntuaciones mayores para valores más altos de calidad, pero en comparaciones *non-target* no se aprecia tal tendencia. Este hecho parece indicar que el poder discriminativo del sistema aumenta con la calidad. Esta observación parece lógica: la degradación de las señales se traduce en una pérdida de información biométrica [Alonso, 2008b] es más probable que dicha degradación haga que dos locuciones de un mismo individuo parezcan diferentes, a que dos locuciones de individuos diferentes se asemejen, o se diferencien más aún.

• Gráficas DET

A continuación se muestran las curvas DET para los sistemas y bases de datos estudiados. Para poder observar la utilidad de las medidas de calidad, se han representado dos curvas para cada sistema y base de datos: la curva original, y la curva correspondiente a la exclusión del 25% de las pruebas con peor calidad. De este modo, se espera que esta segunda curva se sitúe más cerca del origen de coordenadas, lo cual implica una tasa de error menor, y demuestra la utilidad de la medida de calidad.

Las gráficas de la izquierda contienen los resultados para la base de datos SRE 2006, mientras que las de la derecha las de SRE 2008. En cada gráfica se dibujan seis curvas (dos por cada sistema).

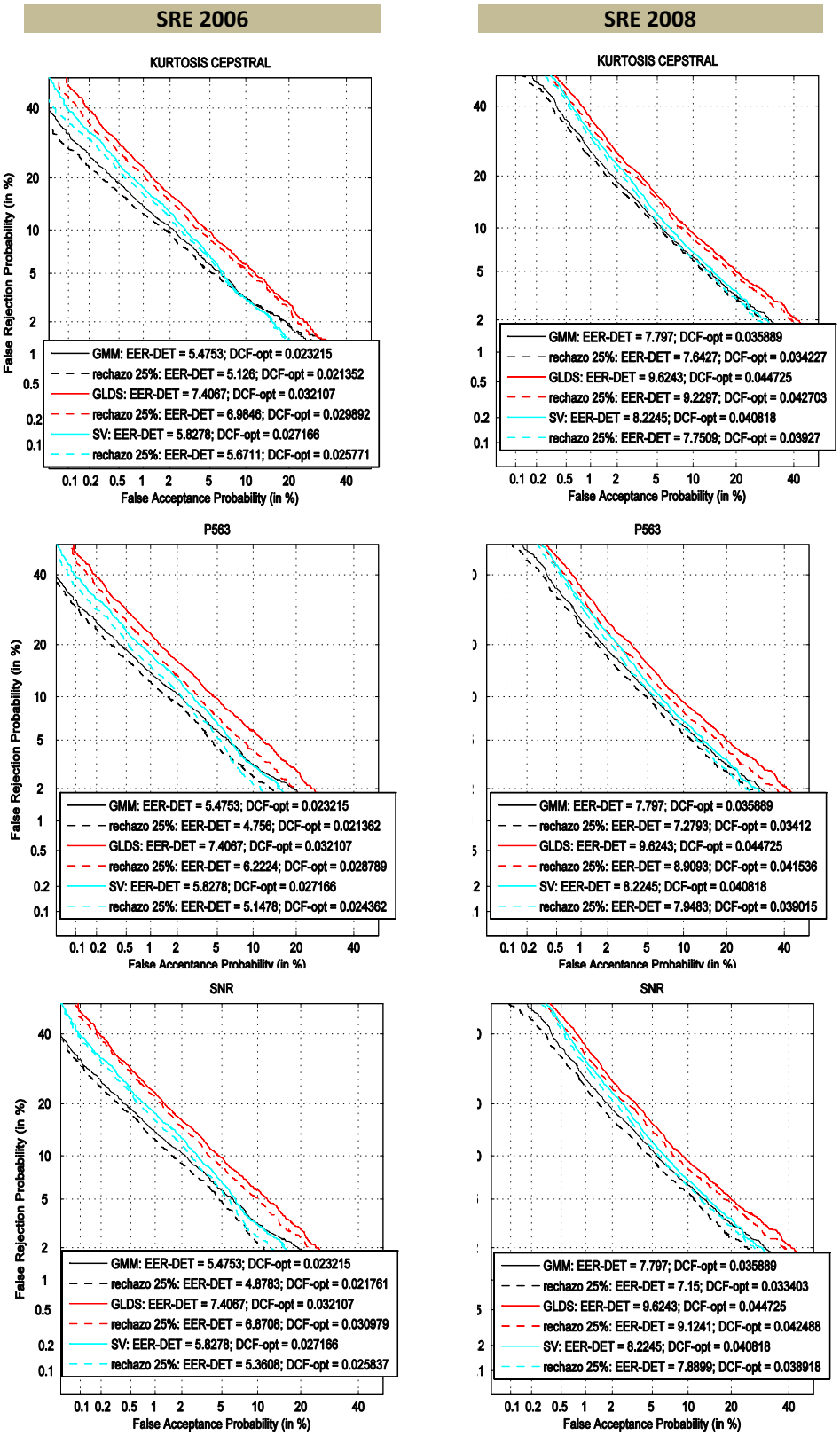


Fig 5.5.a*: Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

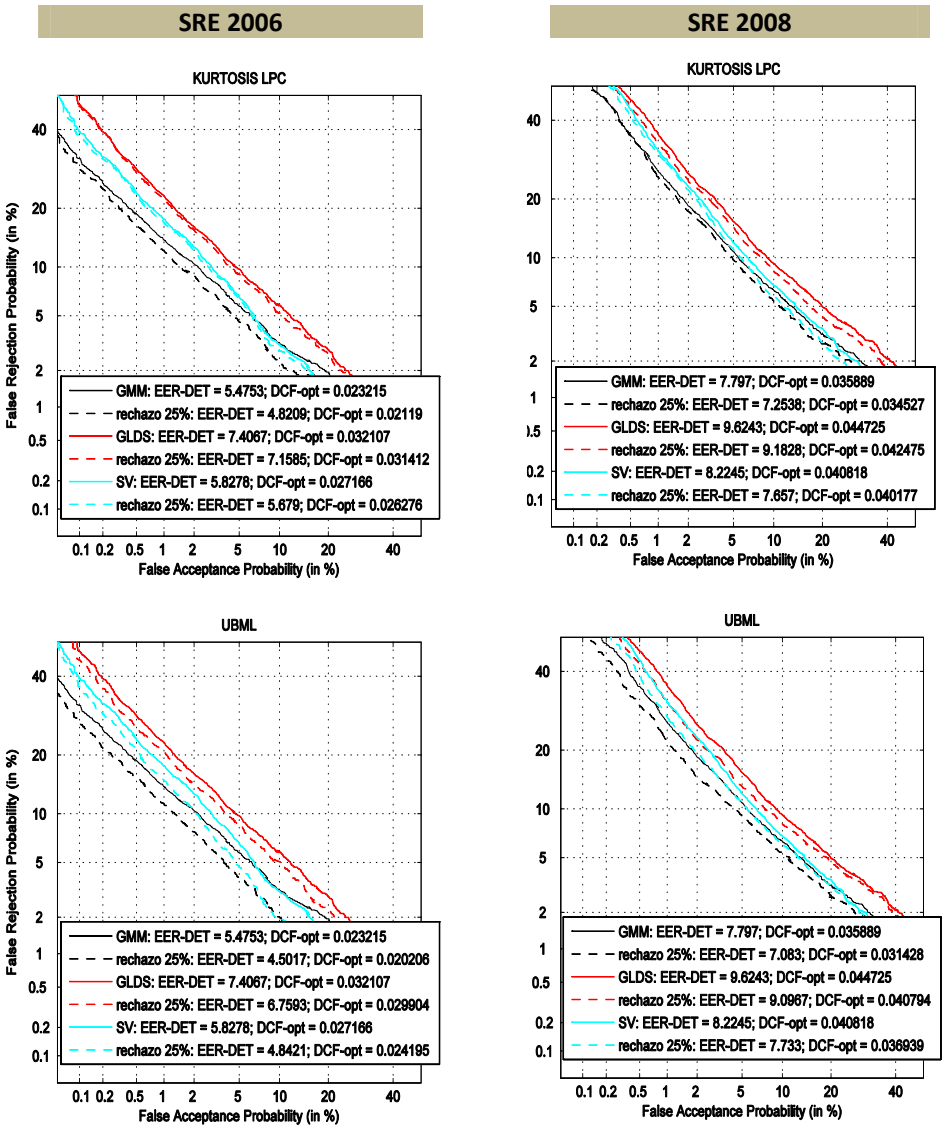


Fig 5.5.b: *Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

En todas las curvas se observa una notable mejora del EER después del rechazo de los scores de peor calidad. Para facilitar el análisis, en la siguiente tabla se reúnen los valores de EER observados. En la primera fila se recogen los valores de EER sin rechazo, y en las sucesivas con el rechazo del 25% de scores con peor calidad.

Sistema / Indicador	SRE 2006			SRE 2008		
	GLDS	SV	GMM	GLDS	SV	GMM
SIN RECHAZO	**7,40%**	**5,83%**	**5,47%**	**9,62%**	**8,22%**	**7,80%**
P563 – 25%	6,22%	5,14%	4,75%	8,9%	7,94%	7,27%
SNR– 25%	6,67%	5,36%	4,87%	9,12%	7,89%	7,19%
KURTOSIS LPC– 25%	7,16%	5,67%	4,82%	9,18%	7,66%	7,25%
KURTOSIS CEPSTRAL– 25%	6,96%	5,57%	5,13%	9,22%	7,75%	7,64%
UBML– 25%	6,75%	4,82%	4,5%	9,09%	7,73%	7,06%

Tabla 5.4: *valores de EER en % para las bases de datos SRE 2006 y SRE 2008, para los sistemas GMM, GLDS y SV.*

Se observa que por lo general las mejoras son mayores para la base de datos SRE 2006. También vemos que para el sistema GMM las mejoras son mayores que para el resto de sistemas. Cabe resaltar que estos son la base de datos y el sistema con los que se consigue mayor poder discriminativo.

En la siguiente tabla se muestra la mejora el EER (en %). De este modo se observa más fácilmente las mejoras experimentadas en cada sistema y base de datos.

Sistema / Indicador	SRE 2006			SRE 2008		
	GLDS	SV	GMM	GLDS	SV	GMM
P563 - 25%	15,95%	11,84%	13,16%	7,48%	3,41%	6,79%
SNR - 25%	9,86%	8,06%	10,97%	5,20%	4,01%	7,82%
KURTOSIS LPC - 25%	3,24%	2,74%	11,88%	4,57%	6,81%	7,05%
KURTOSIS CEPSTRAL - 25%	5,95%	4,46%	6,22%	4,16%	5,72%	2,05%
UBML - 25%	8,78%	17,32%	17,73%	5,51%	5,96%	9,49%
MEDIA	**8,76%**	**8,89%**	**11,99%**	**5,38%**	**5,18%**	**6,64%**

Tabla 5.5: *mejoras del EER en % para las bases de datos SRE 2006 y SRE 2008, para los sistemas GMM, GLDS y SV*

Se confirman las conclusiones que se han comentado. Destaca el comportamiento de la medida propuesta UBML, siendo una de las más útiles y estables de las analizadas.

• Gráficas EER vs Exclusión

A continuación se muestran las figuras "EER vs Exclusión", definidas en la sección 4.2.5, para las dos bases de datos bajo estudio. Se han establecido cinco fracciones de exclusión: 5%,10%, 15%, 20% y 25%.

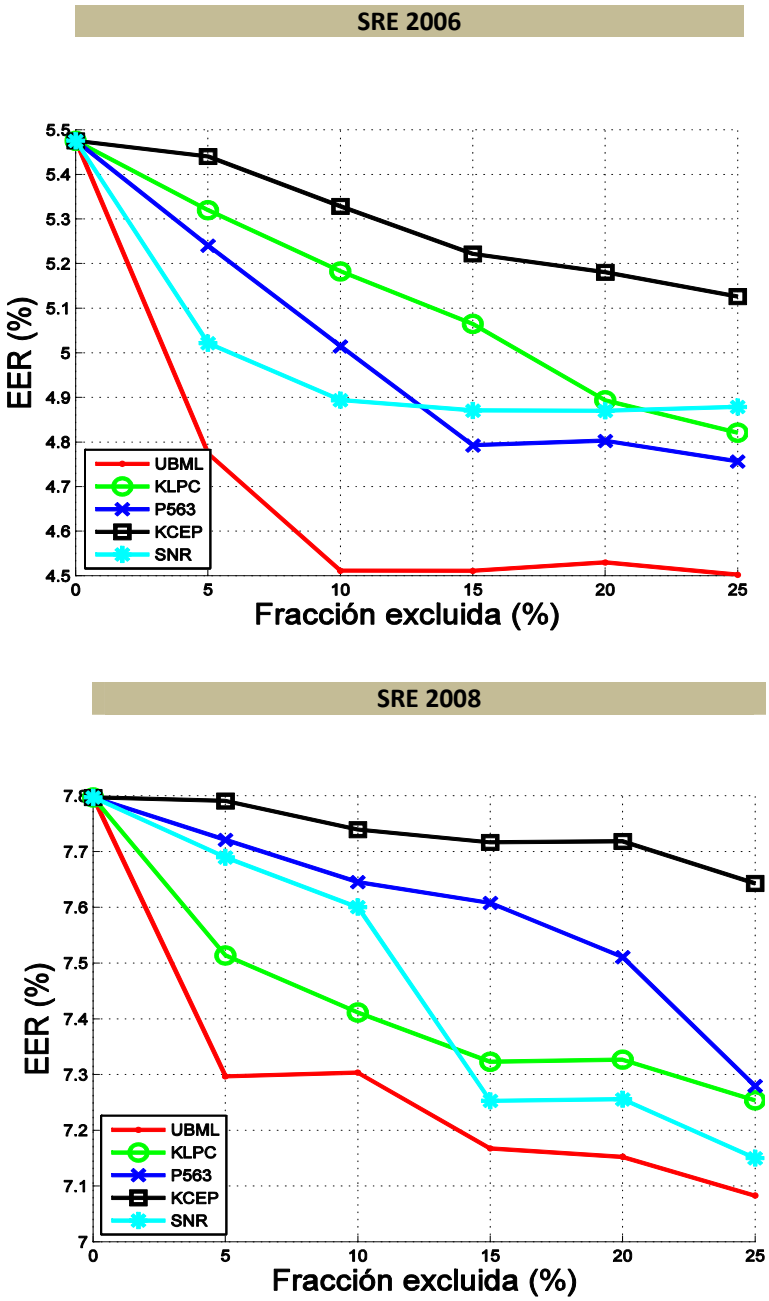


Fig 5.6*: figuras EER (%) frente a fracción excluida de scores para las bases de datos SRE 2006 y 2008.*

Por lo general se observa una tendencia decreciente según aumenta la fracción de scores excluidos. Para algunas medidas, en algunos tramos varia ligeramente la tendencia manteniéndose el EER, pero en ningún caso observamos subidas significativas. Esto indica que efectivamente las medidas de calidad son en mayor o menor medida como predictoras del impacto de la calidad en el sistema.

Observamos que las medidas SNR, P563 y UBML, para la base de datos SRE 2006, por lo general experimentan unas bajadas del EER muy bruscas en las dos primeras exclusiones, para luego mantenerse más o menos estables. Si observamos las gráficas del estudio de IDs, también se observa una caída más brusca en valores más bajos de calidad, lo cual concuerda con esta explicación. Esto indica que un mapeo más ajustado hubiese sido más apropiado para estas medidas, pero decidimos mantener el mapeo lineal para evitar riesgos de sobreajuste al cambiar a otras bases de datos.

Observamos que la medida de calidad propuesta UBML presenta el EER más bajo en todos los tramos en ambas bases de datos. Para esta misma medida se observa el efecto comentado anteriormente, notándose en ambos casos un descenso del EER de más del 50% en la primera exclusión de scores. Esto podría explicarse observando la distribución de dicha medida (ver Anexo A). En este observamos que existen un pequeño conjunto de locuciones que se aleja mucho de la media de la distribución, por lo que en la primera exclusión probablemente se está eliminando gran parte de esa cola, que se corresponde con scores con una calidad muy baja.

5.2 Experimentos con habla microfónica

En esta sección se muestran los resultados obtenidos para los experimentos con bases de datos microfónicas y cruces de habla telefónica y microfónica, que incluyen las siguientes condiciones de la base de datos SRE 2008:

- Micrófono-micrófono (*mic-mic*): tanto las locuciones de testeo como las utilizadas en el entrenamiento de modelos, son de origen microfónico.
- Micrófono-teléfono (*mic-tlf*): modelos entrenados con habla microfónica y locuciones de testeo telefónicas.
- Teléfono-micrófono (*tlf-mic*): modelos entrenados con habla telefónica y locuciones de testeo microfónicas.

Recordemos que existen dos tipos de locuciones microfónicas en la base de datos que utilizamos (sección 4.1.1):

- Locuciones estándar: conversacionales grabadas con diferentes micrófonos de un hablante en una conversación telefónica.
- Locuciones "Interview": el audio contiene la voz del locutor, más la de un entrevistador que realiza las preguntas. La grabación está tomada con diferentes micrófonos.

Las locuciones tipo "Interview" presentan un problema de cara a extraer sus indicadores de degradación, ya que el detector de actividad de voz puede tener dificultades para separar la voz de los silencios, debido a la voz del entrevistador. En los experimentos de correlación que se presentan más adelante pueden observarse estos efectos. Otro inconveniente de este tipo de locuciones es que para cada uno de los sistemas, la voz del entrevistador es eliminada antes de extraer las características de la voz. Sin embargo las medidas de calidad se obtuvieron a partir del fichero entero. Estos detalles deberán ser tenidos en cuenta en la fase de análisis.

Se han estudiado los mismos indicadores de degradación que para los expeimentos con locuciones telefónicas, excepto el indicador SNR por filtrado de Wiener, que no se pudo probar, debido a que el filtro no funcionaba correctamente en las locuciones interview, pues se basa en un detector de actividad de voz para poder estimar el nivel de ruido en los silencios. Como en este tipo de locuciones los silencios se confunden con los fragmentos ocupados por la voz del entrevistador, el filtro de Wiener no puede funcionar.

Cabe destacar que el alcance establecido para los experimentos de esta sección es diferente del de los experimentos de condición tlf-tlf, ya que se experimenta con una sola base de datos y descartamos un indicador de degradación. Sin embargo, se cubren otros aspectos que en la sección de experimentos tlf-tlf (sección 5.1) no se tocan, a saber:

• **Influencia de la calidad en cruces de micrófono y teléfono**. Al disponer de la utilidad de todas las medidas (tanto habla microfónica como telefónica), los experimentos con cruces permiten observar la influencia de la calidad de locuciones de testeo y modelos por separado. Por ejemplo, para una medida con una utilidad mucho mayor en habla telefónica que en microfónica, sería interesante observar como varía la utilidad cuando utilizamos este tipo de habla sólo en el modelo, o sólo en la locución de test .

• **Comparación de la utilidad para habla telefónica, y para microfónica**. Observar cómo varía la utilidad de una misma medida para ambos tipos de habla puede dar pistas de los factores que influyen sobre las distintas medidas: factores propios de las redes telefónicas (como reverberaciones, filtrado,etc.) y factores más independientes del tipo de habla.

• **La utilidad de medidas de tipo subjetivo en habla microfónica**. Permitirá conocer si estimadores que han demostrado gran eficacia en la estimación subjetiva de la calidad de la voz telefónica, como la recomendación P563 o la SNR, pueden valer también para habla microfónica.

Dado que el enfoque para estos experimentos difiere de los anteriores, la estructura también lo hace; de este modo, las secciones de utilidad para cada tipo de experimentos no se presentan hasta el final, donde se realizará un análisis comparativo de todos los experimentos vistos hasta el momento. Así mismo se introduciran algunos subapartados en los que se hagan comparativas con los experimentos de habla telefónica.

5.2.1 Experimentos condición *mic-mic*

Estos experimentos se llevaron a cabo con las locuciones de habla microfónica de la base de datos NIST SRE 2008 (sección 4.1.1).

5.2.1.1 Estudio de indicadores de degradación

A continuación se muestran las gráficas obtenidas para los distintos indicadores de degradación.

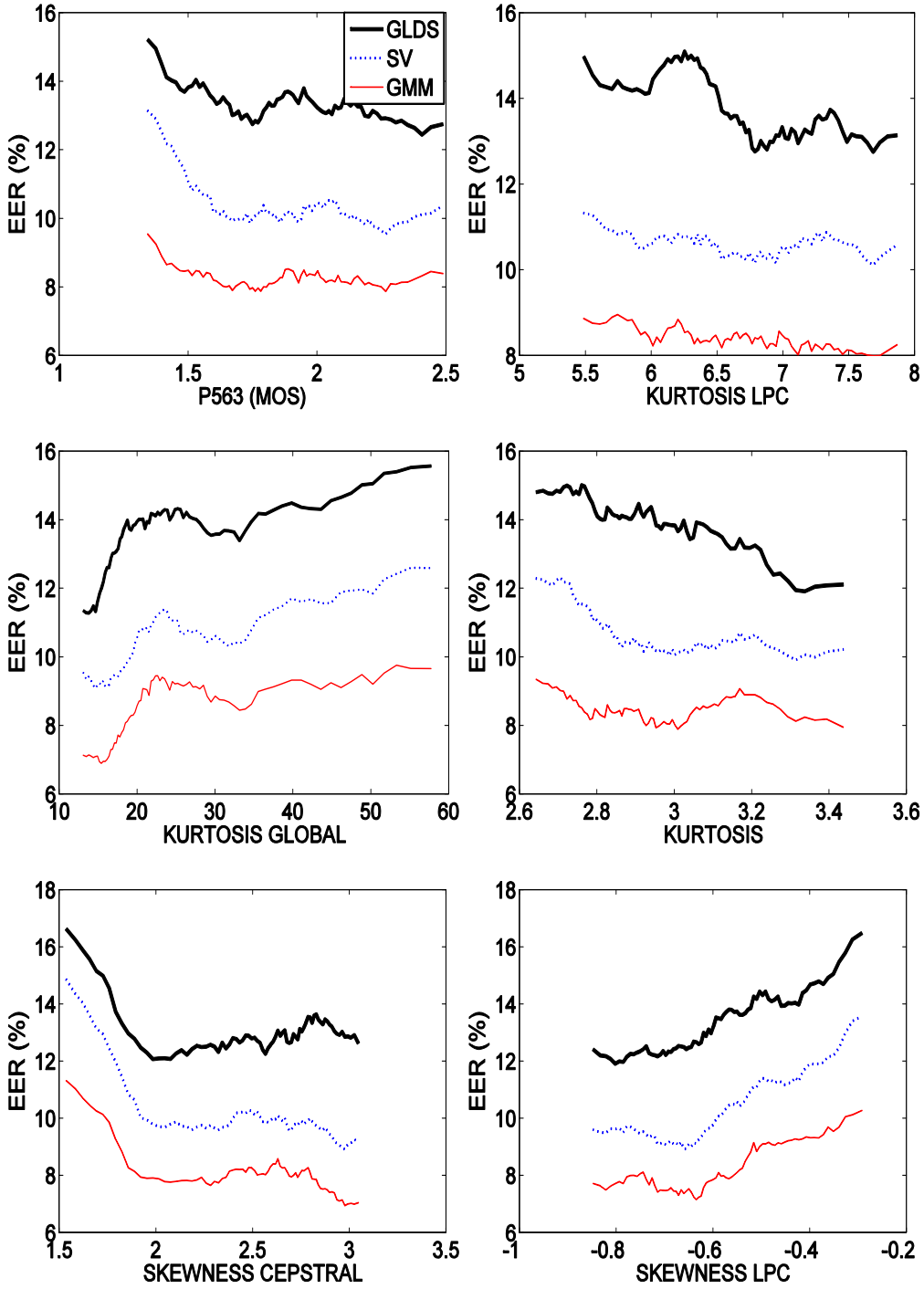


Fig. 5.7.a: *curvas de rendimiento vs Magnitud, para los IDs indicados. Base de datos SRE 2008*

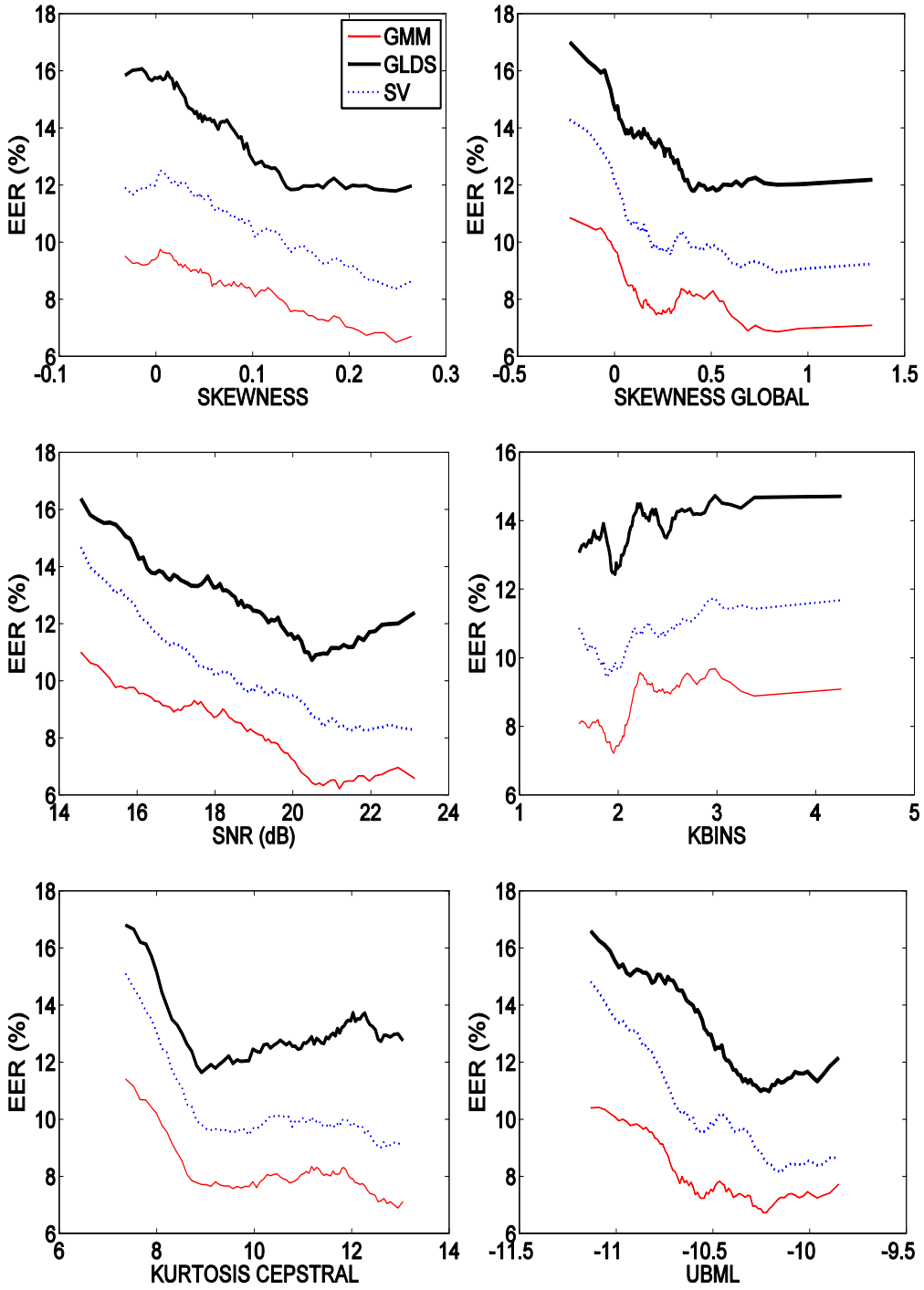


Fig. 5.7.b: *curvas de rendimiento vs Magnitud, para los IDs indicados. Base de datos SRE 2008*

Las conclusiones extraídas de las gráficas se resumen en los siguientes puntos:

• **Relación entre el rendimiento y la magnitud del indicador (impacto**). Por lo general se observa una clara relación entre los IDs y el rendimiento de los sistemas. En la siguiente tabla se muestran las mejoras en % para cada indicador, calculada como el porcentaje que

representa la diferencia entre el mayor y el menor EER observado con respecto al máximo EER.

Sistema / Indicador	SRE 2008			Media
	GLDS	SV	GMM	
P563	18%	27%	18%	**21%**
SNR	35%	44%	44%	**41%**
KURTOSIS	21%	19%	16%	**19%**
SKEWNESS	27%	33%	33%	**31%**
KURTOSIS LPC	16%	11%	11%	**13%**
SKEWNESS LPC	28%	34%	30%	**31%**
KURTOSIS CEPSTRAL	31%	40%	40%	**37%**
SKEWNESS CEPSTRAL	40%	27%	39%	**36%**
KURTOSIS GLOBAL	28%	29%	31%	**29%**
SKEWNESS GLOBAL	37%	31%	37%	**35%**
KBINS	16%	20%	25%	**20%**
UBML	34%	45%	35%	**38%**
Media	**28%**	**30%**	**30%**	**29%**

Tabla 5.6: *mejoras del EER (%) para la condición mic-mic de SRE 2008*.

Observamos que la SNR y UBML (que experimentaban un impacto alto con bases de datos telefónicas), son los IDs que tienen un mayor impacto. Los IDs basados en coeficientes cepstrales (KCEP y SCEP) muestran un impacto muy próximo a estos. Más adelante se comparan estos valores con los que se obtuvieron en habla telefónica.

• **Coherencia entre sistemas.** Se puede observar que la tendencia de las curvas es muy similar entre los tres sistemas para todos los IDs, por lo que el comportamiento es coherente entre sistemas. Además, cabe destacar que la tendencia es bastante más clara que en las gráficas de locuciones telefónicas, observándose oscilaciones mucho menos bruscas. Este hecho podría ser explicado por la variabilidad de micrófonos con la que se registraron las locuciones, ya que dichos micrófonos eran de calidades muy dispares, lo cual se traduce en locuciones de calidades diferentes, lo cual podría estar produciendo una tendencia tan marcada en las gráficas.

• **Comparación con los IDs en habla telefónica**

La comparación con los IDs en habla telefónica se divide en dos puntos: la tendencia de las curvas (creciente o decreciente) y el impacto (para compararlos cuantitativamente).

• **Tendencia**

Para realizar una comparativa de la tendencia entre ambas bases de datos, se ha creado la siguiente tabla, en la cual "↓" indica una tendencia decreciente, "↑" una tendencia creciente, y "---", sin una tendencia clara.

	Tendencia	
BBDD / Indicador	Telf	Mic
P563	↓	↓
SNR	↓	↓
SNR-WIENER	---	N/A
KURTOSIS LOCAL	↓	↓
SKEWNESS LOCAL	---	↓
KURTOSIS LPC	**↑**	**↓**
SKEWNESS LPC	**↓**	**↑**
KURTOSIS CEPSTRAL	↓	↓
SKEWNESS CEPSTRAL	↓	↓
KURTOSIS GLOBAL	---	↑
SKEWNESS GLOBAL	---	↓
KBINS	---	---
UBML	↓	↓

Tabla 5.7*: tendencia de las curvas de estudio de los IDs para las condiciones telefónica (izquierda) y microfónica (derecha).*

Cabe destacar que hay dos indicadores que cambian la tendencia muy significativamente: KLPC y SLPC, cuyas curvas pasan de crecer a decrecer y viceversa.

En otros IDs, se observa una tendencia clara en el habla microfónica, mientras que en la telefónica no se apreciaba tal tendencia (SKEWNESS GLOBAL, SKEWNESS LOCAL,KURTOSIS GLOBAL).

Para el resto de indicadores se mantiene la tendencia observada en habla telefónica.

• **Impacto**

En la siguiente tabla se recogen los valores medios de mejora del EER (%) para ambos tipos de locuciones.

Tipo / Indicador	Teléfono	Micrófono
P563	**44%**	**21%**
SNR	**32%**	**41%**
KURTOSIS	**30%**	**19%**
SKEWNESS	**24%**	**31%**
KURTOSIS LPC	**36%**	**13%**
SKEWNESS LPC	**31%**	**31%**
KURTOSIS CEPSTRAL	**34%**	**37%**
SKEWNESS CEPSTRAL	**33%**	**35%**
KURTOSIS GLOBAL	**29%**	**29%**
SKEWNESS GLOBAL	**17%**	**35%**
KBINS	**29%**	**20%**
UBML	**44%**	**38%**
Media	**32%**	**29%**

Tabla 5.8: *mejoras medias del EER (%) para experimentos telefónicos y microfónicos.*

De los tres IDs con mayor impacto en teléfono, el P563 es el único que disminuye considerablemente. Probablemente esto se debe a que el estimador de calidad P563 está diseñado para funcionar con habla telefónica, que suele estar degradada factores propios de este tipo de habla (ecos, filtrado, diafonía, etc.) que no afectan al habla microfónica, y por tanto disminuye su precisión con esta última.

El indicador SNR mantiene una mejora alta, siendo incluso más alta que en el caso del habla telefónica. El valor medio de SNR para las locuciones microfónicas es 20,2 dB (14 dB más bajo ver Anexo A), por lo que parece coherente que esta mejora se mantenga alta, al igual que podría ser una causa de que el rendimiento para los tres sistemas disminuya con respecto a los resultados de habla telefónica.

Por otro lado cabe destacar que los valores del ID SLPC no implican que el impacto sea similar, ya que su tendencia es contraria, al igual que pasa con el KLPC.

El indicador propuesto UBML sigue manteniendo un alto impacto en ambos tipos de habla, al igual que mantiene una tendencia clara y coherente entre los tres sistemas.

5.2.1.2 Experimentos de correlación

Siguiendo el mismo procedimiento que en experimentos con habla telefónica, se han generado los *Scatter-Plots* para los IDs que han demostrado ser más prometedores, con sus correspondientes coeficientes de correlación lineal. En este caso han sido los siguientes:

- P563
- UBML

- SNR
- SLPC
- SCEP
- KCEP
- KURTOSIS

.

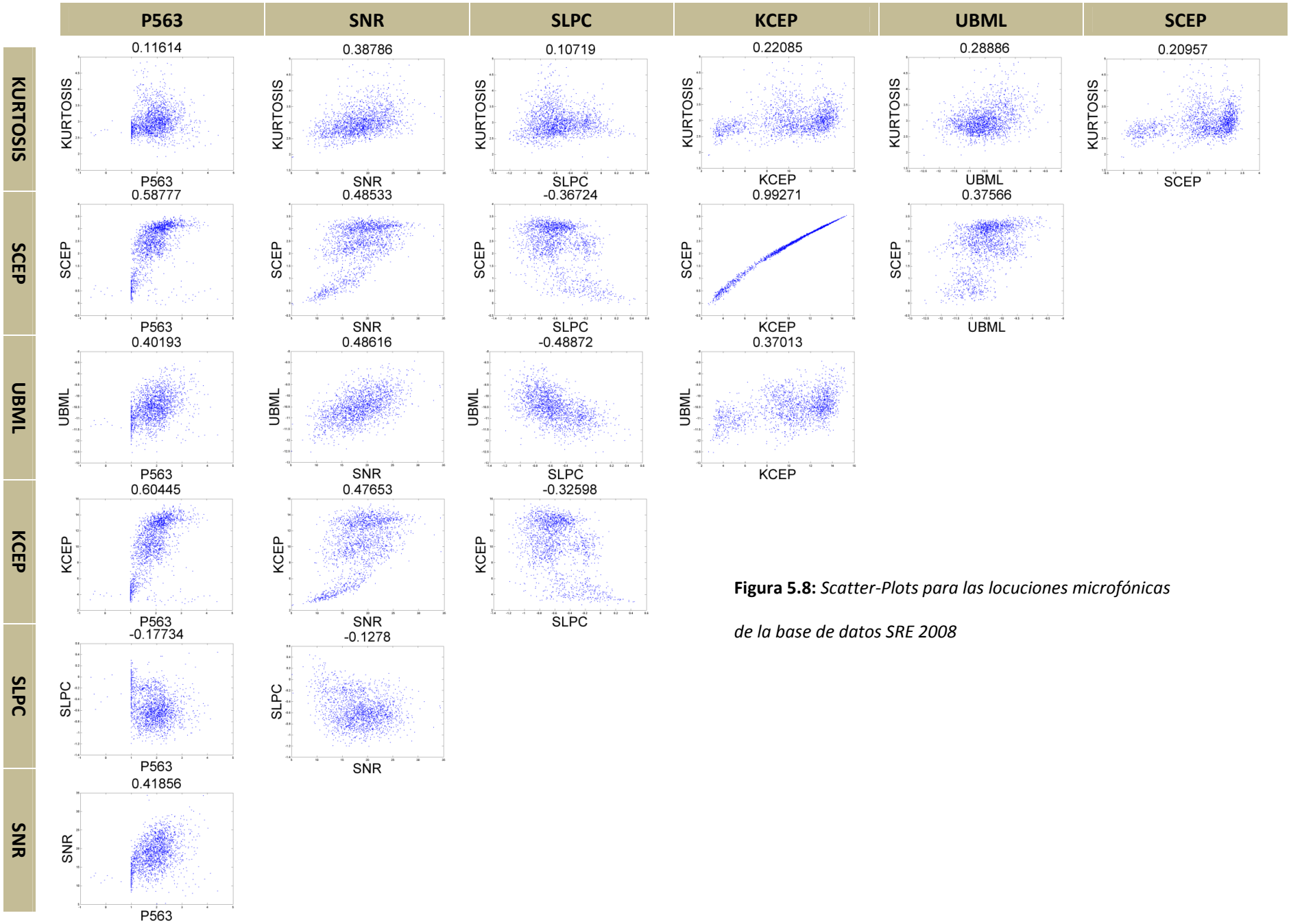


Figura 5.8: *Scatter-Plots para las locuciones microfónicas de la base de datos SRE 2008*

Por lo general se observa que los coeficientes de correlación son notablemente más altos que en los experimentos con locuciones únicamente telefónicas.

En todas las gráficas se observan dos nubes de puntos diferenciadas, debidas a los dos tipos de locuciones mencionadas. Se calculó la media de los IDs para las locuciones estándar y las *interview* por separado, y esta última resultó tener un valor indicativo de una calidad más baja, en general para todas las medidas de calidad analizadas, lo cual parece indicar que la nube de los IDs de *interview* puntúa en valores que indican una calidad inferior a las locuciones estándar. Esto no implica necesariamente que sean locuciones de una calidad inferior; podría ser que la voz del entrevistador produzca este efecto, ya que todos los IDs se basan en la utilización de un detector de actividad de voz para separar la voz de los silencios. Al existir dos nubes diferenciadas los valores de correlación de los distintos IDs podrían estar viéndose distorsionados, lo que podría explicar los altos valores de correlación.

Por último, cabe resaltar la forma de algunas de las nubes de puntos, que parecen indicar que la relación entre algunas parejas de indicadores es no lineal. Dos ejemplos claros serían KCEP-P563 y la KCEP-SNR.

5.2.1.3 Mapeo

Dado que estamos definiendo las funciones de mapeo con una aproximación lineal, para definir las funciones de mapeo de las medidas de calidad en habla microfónica bastará con determinar si tienen la misma tendencia y los mismos rangos de variación que en habla telefónica. Ver Anexo A para consultar los rangos de variación de los IDs.

Los IDs seleccionados para ser mapeados a medida de calidad, tienen todos la misma tendencia y rango de variación que en habla telefónica, a excepción del indicador SLPC, para el cual no se definió función de mapeo anteriormente. Por lo tanto bastará con definir la de este último.

Indicador	Rango	Aproximación	Función de mapeo
SLPC	(-1.5,0.5)	Lineal decreciente	$Q_{SLPC}(x) = 1 - \frac{(x+1{,}5)}{2}$

Tabla 5.9: *función de mapeo a medida de calidad para el indicador SLPC.*

• Gráficas Score vs Calidad

En el caso de las locuciones microfónicas, sólo se presentará una muestra de dos medidas de calidad, a modo de ejemplo, con el fin de corroborar los efectos observados en las mismas gráficas para experimentos de habla telefónica.

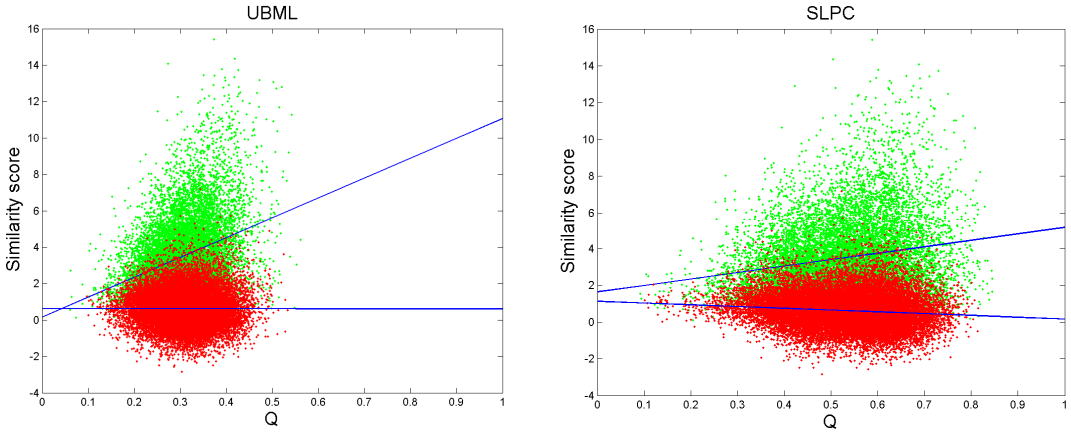


Fig 5.9: *gráficas de dispersión Scores vs Calidad (Q), para los indicadores UBML y SLPC, para la base de datos SRE 2008 (canal microfónico) y sistema GMM.*

Como podemos observar, al igual que en los experimentos telefónicos, las dos nubes tienden a separarse para mayores valores de calidad, siendo notablemente mayor dicho efecto para la nube de comparaciones target.

5.2.2 Cruces micrófono-teléfono

Los resultados que se presentan a continuación corresponden a comparaciones de locuciones de testeo telefónicas con modelos generados a partir de habla microfónica.

5.2.2.1 Estudio de indicadores de degradación

A continuación se muestran las gráficas "Rendimiento vs Magnitud" para los IDs estudiados, que han sido generadas por el mismo procedimiento seguido anteriormente.

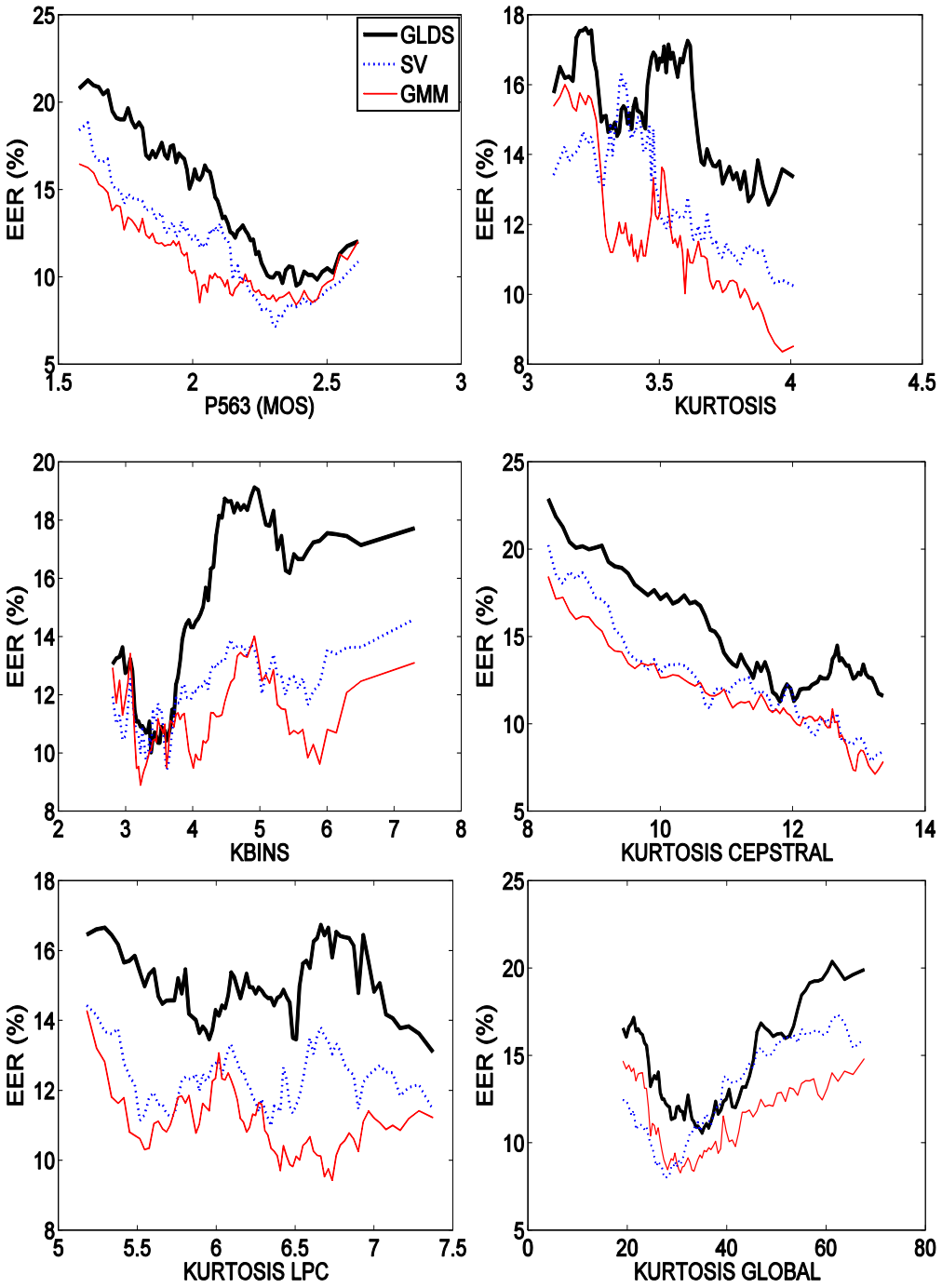


Fig 5.10.a *Curvas Rendimiento vs Magnitud para los IDs indicados bajo las gráficas, en la condición mic-tlf de SRE 2008.*

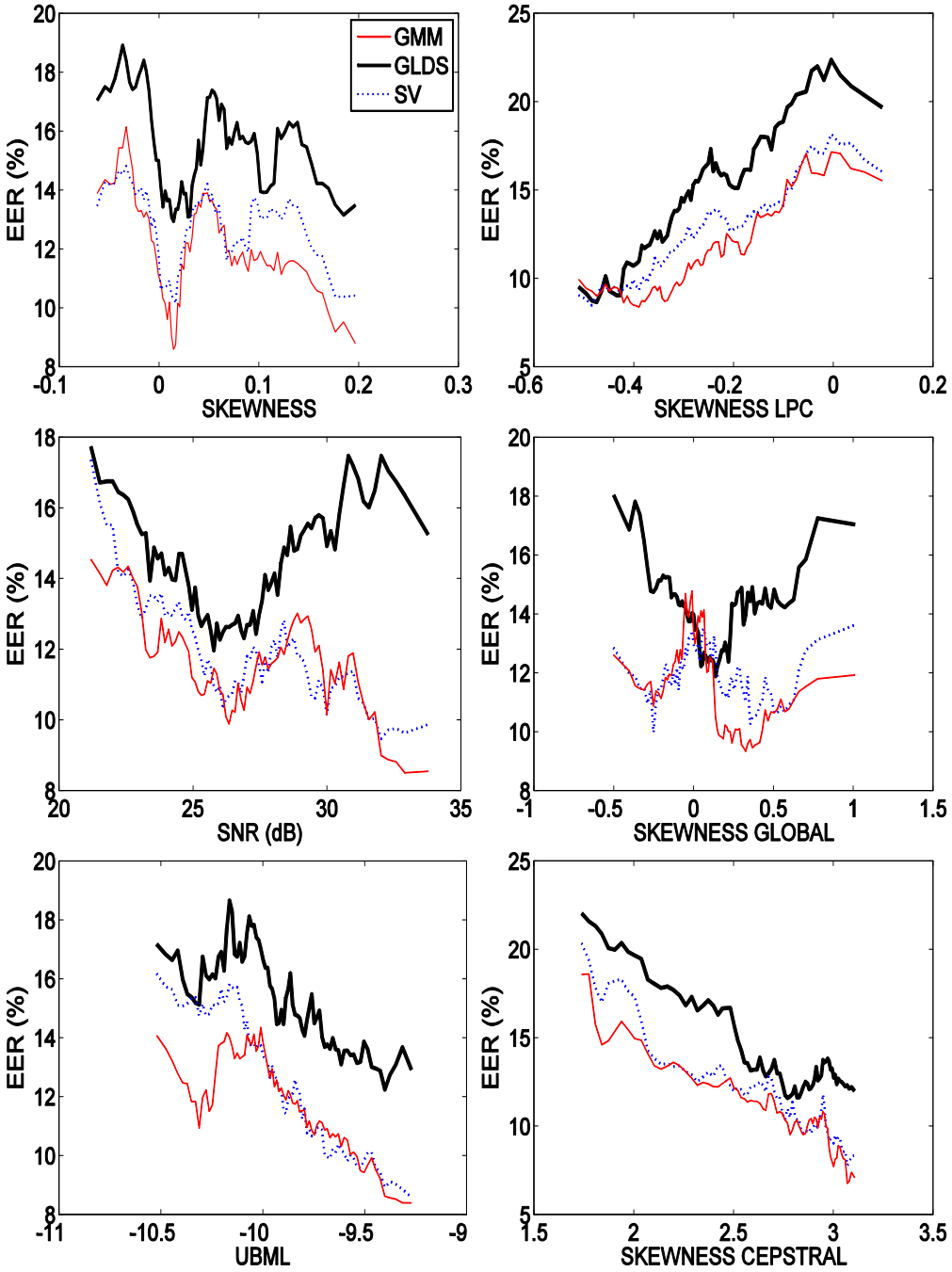


Fig 5.10.b: *Curvas Rendimiento vs Magnitud para los IDs indicados bajo las gráficas, en la condición mic-tlf de SRE 2008.*

Las conclusiones extraídas de las gráficas, se dividen como anteriormente en dos puntos:

• **Impacto.** Se observa que los IDs que habían demostrado mayor utilidad en ambos tipos de habla siguen teniendo un impacto importante en todos los sistemas (SNR,P563 y UBML).

En la siguiente tabla se recogen las máximas mejoras del EER para todos los IDs y sistemas**:**

Sistema / Indicador	GLDS	SV	GMM	Media
P563	55%	62%	49%	**55%**
SNR	41%	45%	32%	**39%**
KURTOSIS	29%	37%	48%	**38%**
SKEWNESS	32%	32%	47%	**37%**
KURTOSIS LPC	22%	24%	34%	**27%**
SKEWNESS LPC	61%	53%	51%	**55%**
KURTOSIS CEPSTRAL	61%	51%	61%	**58%**
SKEWNESS CEPSTRAL	62%	47%	63%	**57%**
KURTOSIS GLOBAL	48%	51%	43%	**47%**
SKEWNESS GLOBAL	34%	27%	37%	**33%**
KBINS	48%	35%	37%	**40%**
UBML	34%	47%	42%	**41%**
Media	**44%**	**43%**	**45%**	**44%**

Tabla 5.10: *mejoras del EER (%) en el estudio de los IDs para cruces micrófono-teléfono. Base de datos SRE 2008*

Observamos que destacan dos IDs por sus valores: KCEP y SCEP, que tienen un valor notablemente mayor que el experimentado en habla telefónica y microfónica. Esto mismo ocurre con la SLPC, y en el resto de IDs, en mayor o menor medida, también sucede. Esto puede ser debido a que la tasa de error (EER) de estos experimentos es mayor que en los anteriores (debido al desajuste de bases de datos de distinto tipo de habla), y de este modo la información sobre la calidad podría estar ayudando a compensar este desajuste en mayor medida de lo que lo hace en los anteriores experimentos.

• **Coherencia entre sistemas.** Se observa que por lo general el comportamiento de las curvas es coherente en IDs que ya habían demostrado ser útiles tanto en experimentos telefónicos como en microfónicos, mientras que en aquellos IDs que han demostrado un impacto bajo, el comportamiento de las curvas difiere. Por lo general no hay una coherencia tan clara como en los experimentos de habla únicamente microfónica.

• **Gráficas Score vs Calidad**

Al igual que en el apartado anterior, se muestran estas gráficas a modo de ejemplo.

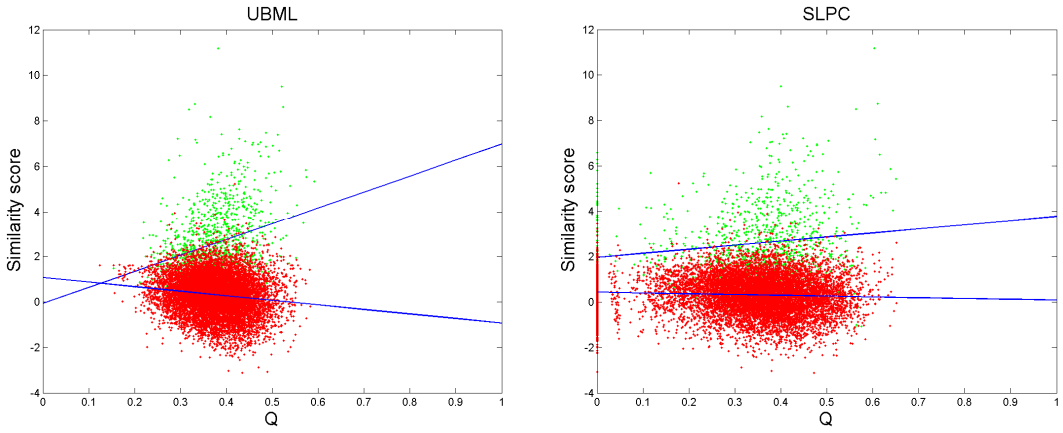


Fig 5.11: *figuras Score vs Calidad para los indicadores UBML y SLPC para la condición mic-tlf de SRE 2008.*

Una vez más se observa que las nubes de usuario e impostor se alejan según incrementa la calidad, siendo este efecto mayor en la nube de comparaciones target.

5.2.3 Cruces teléfono-micrófono

Siguiendo la misma metodología, a continuación se muestran las gráficas obtenidas para los IDs estudiados**.**

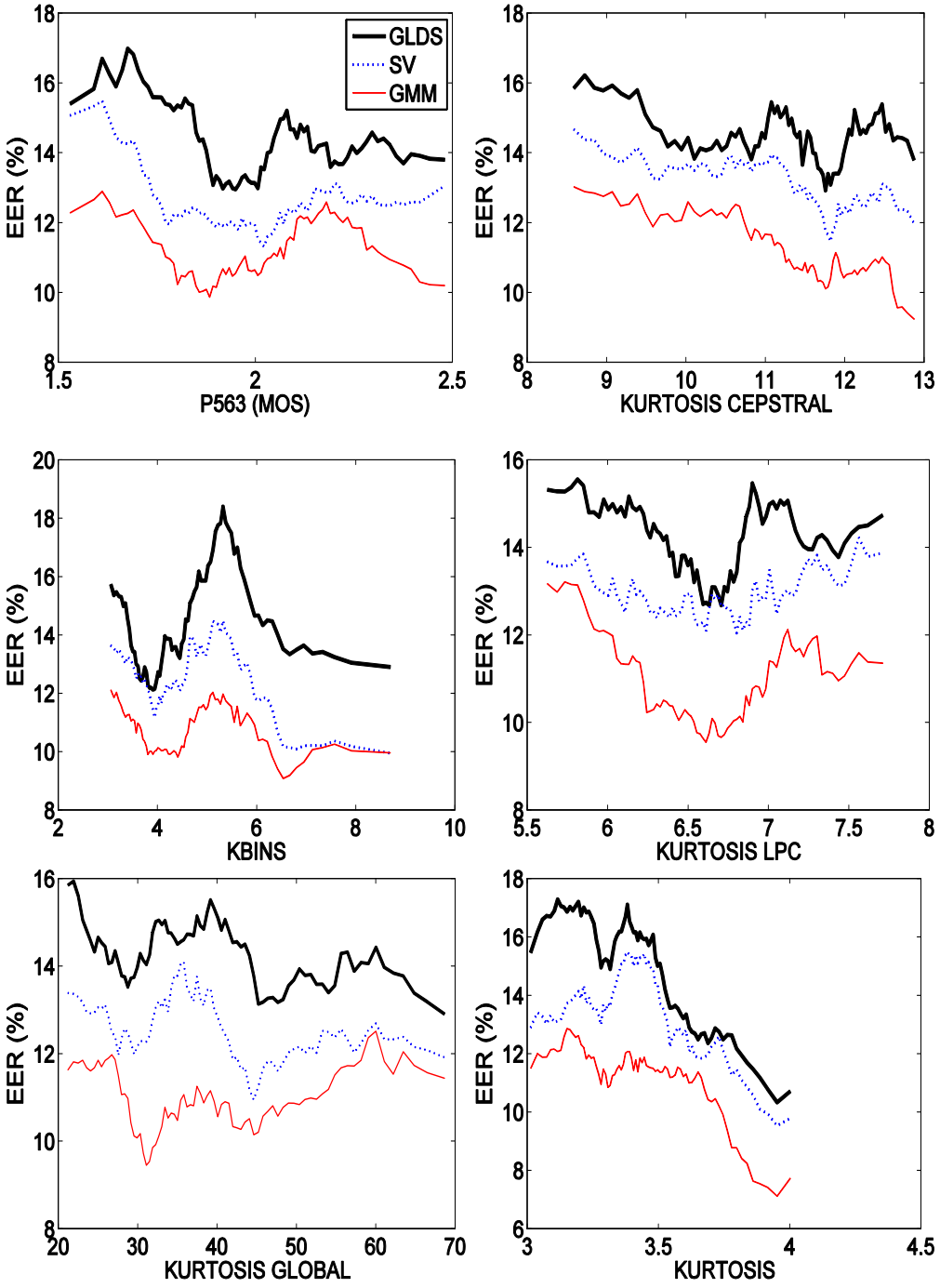


Fig 5.12.a: *Curvas Rendimiento vs Magnitud para los IDs indicados bajo las gráficas, en la condición tlf-mic de SRE 2008.*

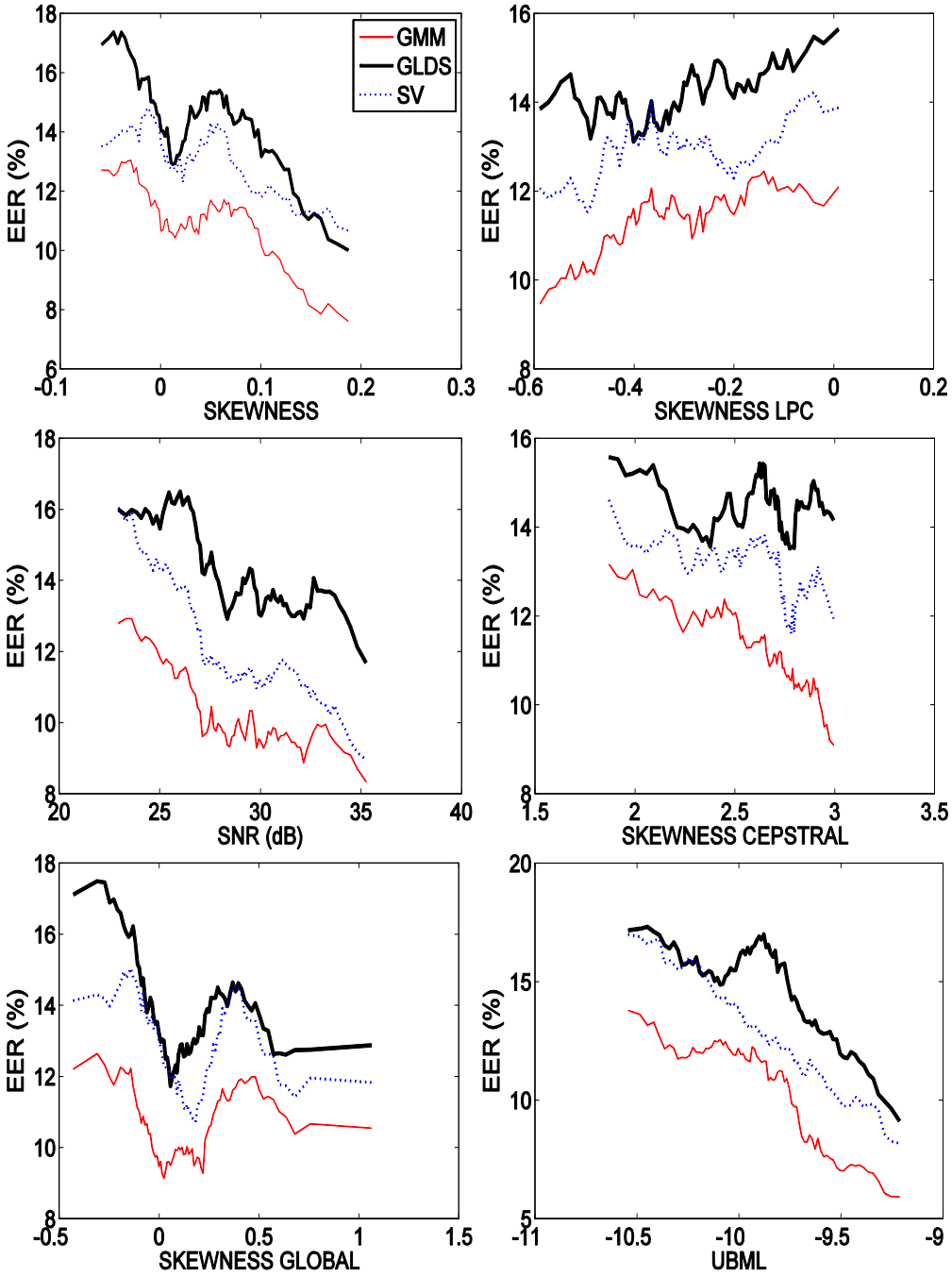


Fig 5.12.b: *Curvas Rendimiento vs Magnitud para los IDs indicados bajo las gráficas, en la condición tlf-mic de SRE 2008.*

Las conclusiones extraídas de las gráficas, se dividen como anteriormente en dos puntos:

•**Impacto.** Observamos que generalmente las pendientes de las curvas son menos pronunciadas que en el caso de cruce micrófono-teléfono, y la relación entre los IDs y el rendimiento de los sistemas es menos clara que en todos los experimentos anteriores, observándose oscilaciones y cambios de pendiente.

En la siguiente tabla se recogen las máximas mejoras del EER para todos los IDs y sistemas**:**

	SRE 2008			Media
Sistema / Indicador	GLDS	SV	GMM	
P563	24%	27%	24%	**25%**
SNR	29%	44%	35%	**36%**
KURTOSIS	40%	38%	45%	**41%**
SKEWNESS	42%	28%	42%	**37%**
KURTOSIS LPC	22%	24%	34%	**27%**
SKEWNESS LPC	19%	15%	28%	**21%**
KURTOSIS CEPSTRAL	20%	22%	29%	**24%**
SKEWNESS CEPSTRAL	13%	21%	31%	**22%**
KURTOSIS GLOBAL	19%	22%	21%	**21%**
SKEWNESS GLOBAL	33%	29%	28%	**30%**
KBINS	34%	31%	25%	**30%**
UBML	47%	52%	57%	**52%**
Media	**29%**	**29%**	**33%**	**30%**

Tabla 5.11: *mejoras del EER (%) en el estudio de los IDs para cruces micrófono-teléfono. Base de datos SRE 2008.*

Como podemos observar, existe una bajada generalizada del impacto. Esta bajada se nota especialmente en el indicador P563, que disminuye su impacto y empeora considerablemente su tendencia con respecto a los cruces micrófono-teléfono. Generalmente la bajada del impacto es más leve en aquellos IDs que demostraron tener un impacto alto tanto en habla microfónica como en habla telefónica (SNR, UBML).

• **Coherencia entre sistemas.** Observamos que, al igual que en los cruces micrófono-teléfono, se observan más oscilaciones que hacen disminuir el grado de paralelismo entre las curvas, aunque podemos considerar que el comportamiento entre sistemas es coherente en todos los casos.

- **Gráficas Score vs Calidad**

Al igual que en el apartado anterior, se muestran estas gráficas a modo de ejemplo.

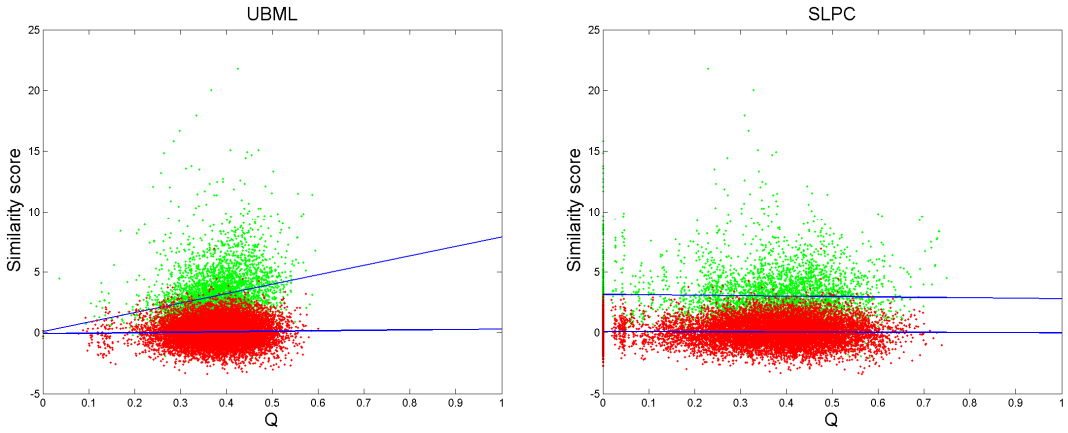


Fig. 5.13: *gráficas de dispersión Scores vs calidad (Q), para las medidas de calidad UBML y SLPC*

Se observa que para la medida de calidad UBML las nubes de usuario e impostor se alejan según incrementa la calidad, siendo este efecto mayor en la nube de comparaciones target. Esto no sucede con la SLPC, algo que era de esperar por los resultados de su estudio como indicador de degradación.

5.2.4 Experimentos de utilidad. Análisis comparativo de la utilidad de las medidas de calidad.

Las gráficas que se presentan a continuación tienen dos objetivos:

•Observar la utilidad de las distintas medidas de calidad con las que se ha experimentado.
•Llevar a cabo un análisis comparativo de las distintas medidas en distintos tipos de condiciones.

Para cada medida de calidad se han dispuesto cuatro gráficas DET en cada línea, correspondientes a las cuatro condiciones experimentales utilizadas hasta el momento (*tlf-tlf, mic-mic, tlf-mic y mic-tlf*), y cada una de las gráficas con seis curvas, correspondientes a los tres sistemas que hemos estudiado. Para cada sistema se representarán la curva original y la resultante de excluir el 25% de la población de scores con peor calidad.

Se han escogido un total de 7 medidas de calidad, que se corresponden con las estudiadas en experimentos tlf-tlf, más las medidas SLPC y Kurtosis Local. El motivo de ampliar el número de medidas es disponer de mayor información a la hora de realizar la comparativa.

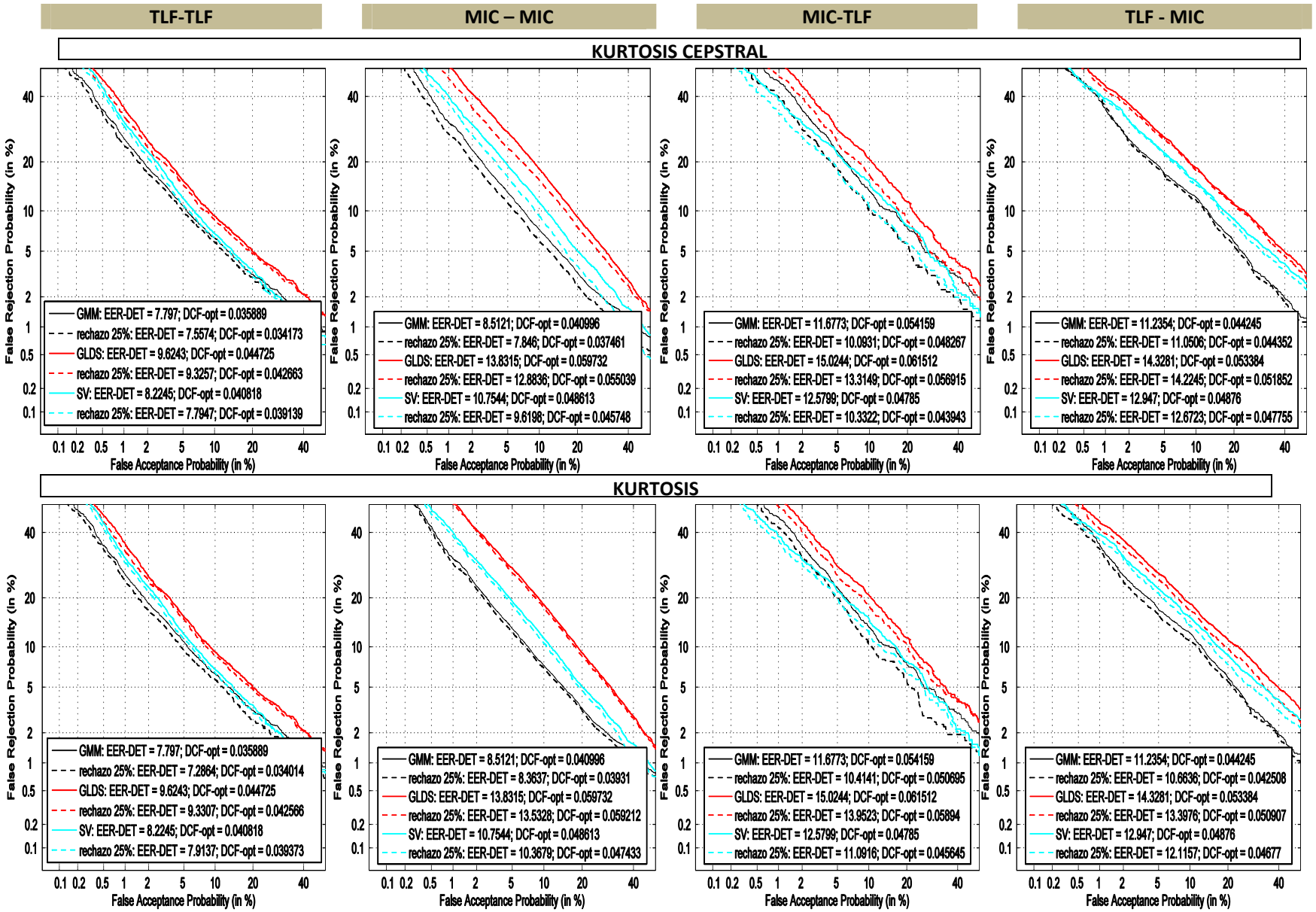


Fig 5.14.a: *Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

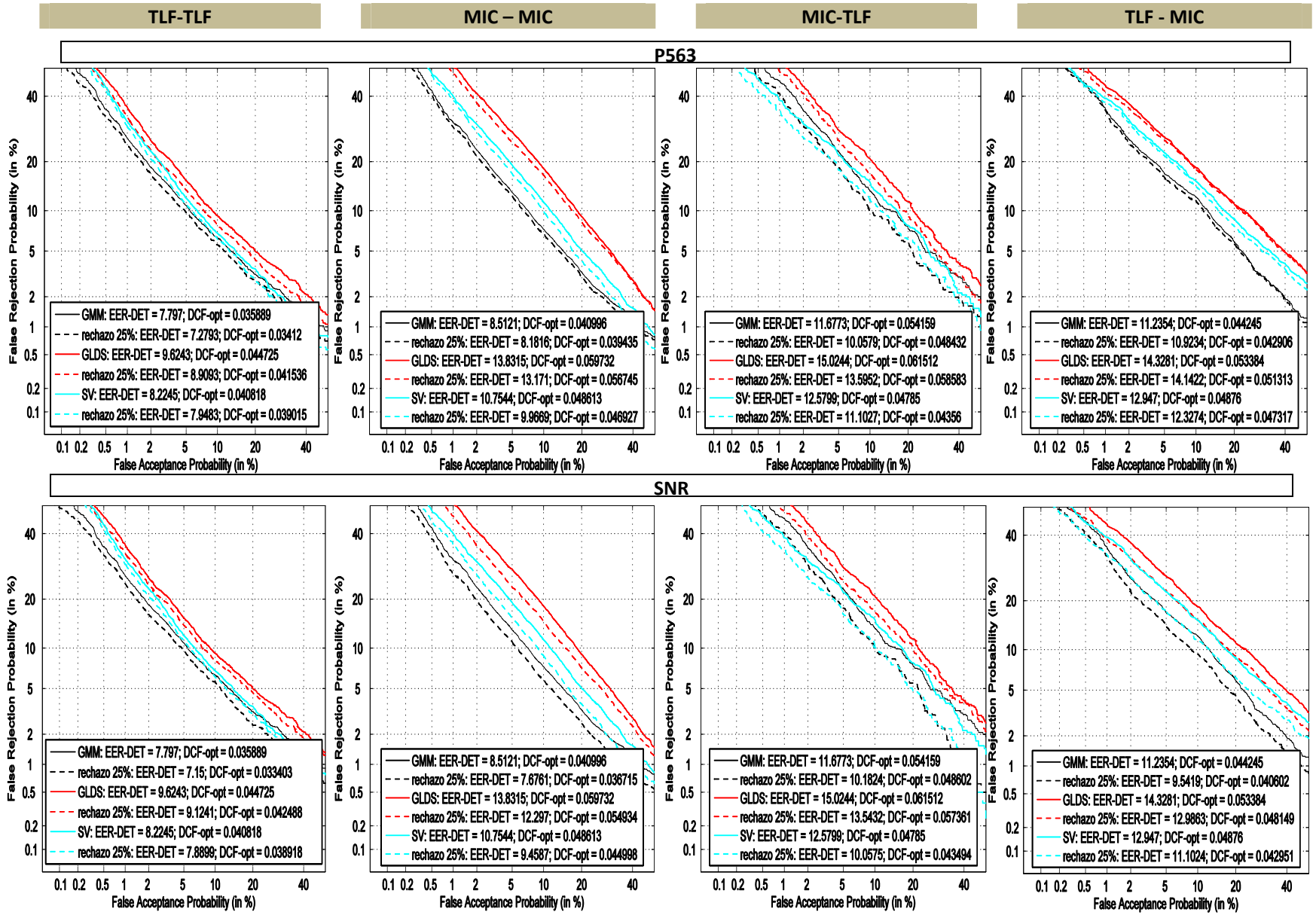


Fig 5.14.b: *Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

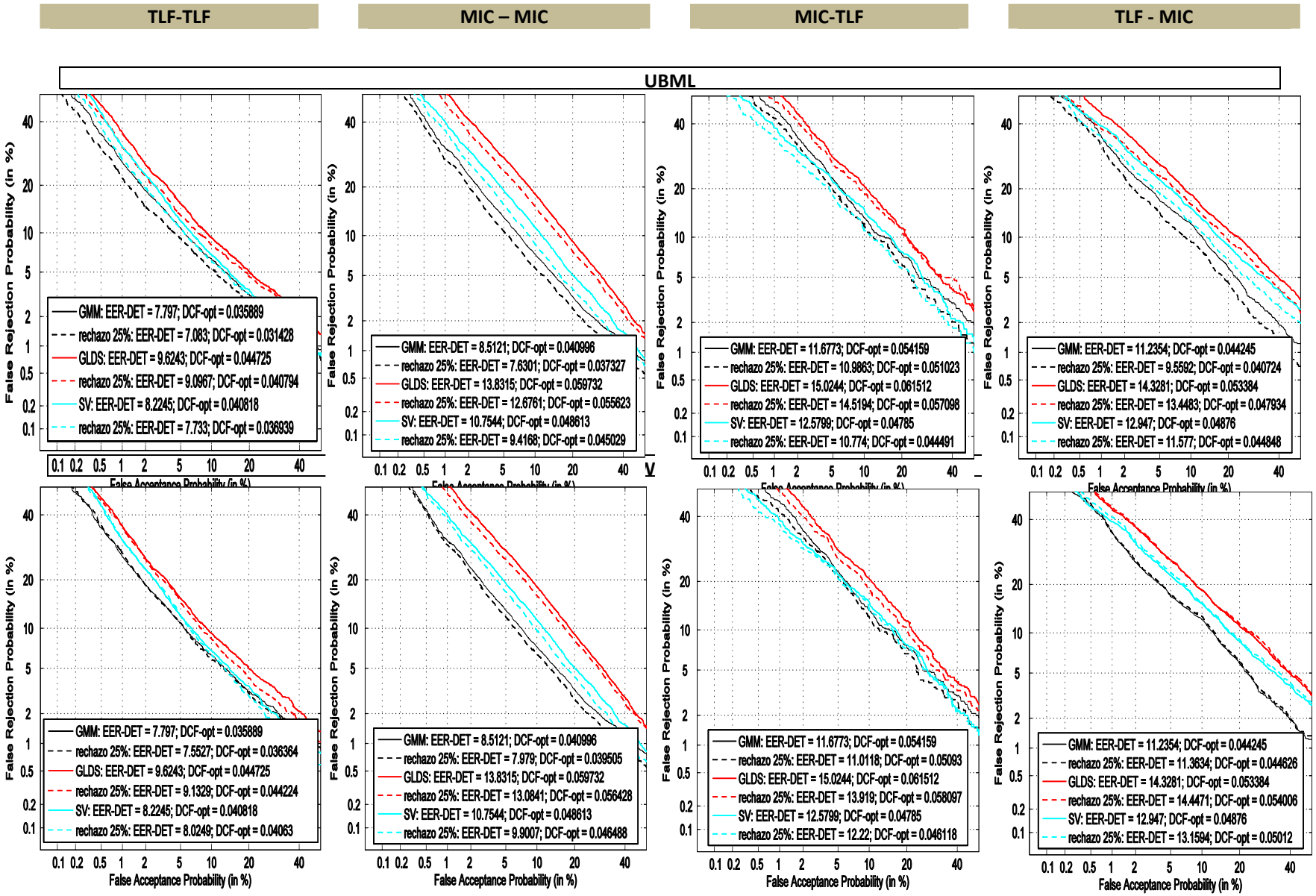


Fig 5.14.c: *Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

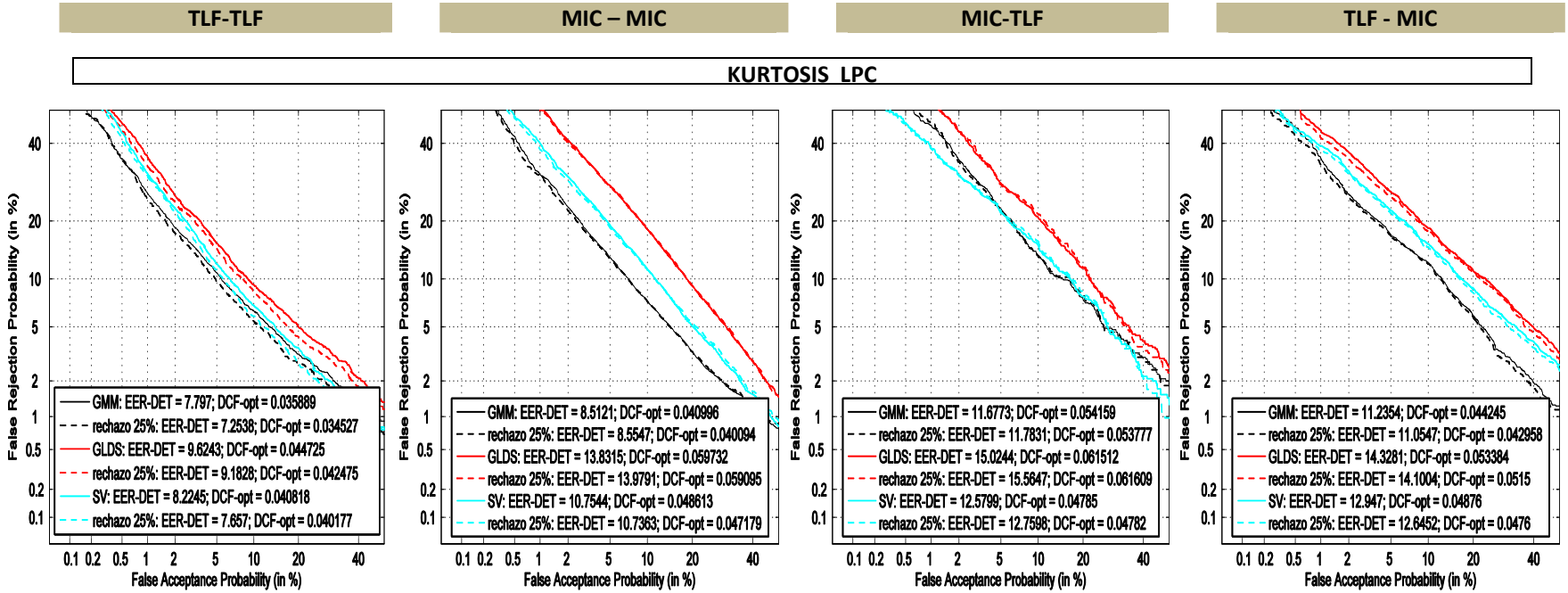


Fig 5.14.d: *Curvas DET para los sistemas GMM,GLDS y SV con 2 curvas por sistema: original y excluidos el 25 % de los scores con calidad más baja, para las bases de datos SRE 2006 (izquierda) y SRE 2008 (derecha).*

	MODELO		TEST	MODELO		TEST	MODELO		TEST	MODELO		TEST
	Tlf		Tlf	Mic		Mic	Mic		Tlf	Tlf		Mic
SISTEMA	GMM	GLDS	SV	GMM	GLDS	SV	GMM	GLDS	SV	GMM	GLDS	SV
SIN EXCLUSIÓN	**7,8 %**	**9,6 %**	**8,2 %**	**8,5%**	**13,8%**	**10,8%**	**11,7%**	**15 %**	**12,6%**	**11,2%**	**14,3%**	**13%**
KCEP – 25%	7,6 %	9,3 %	7,8 %	7,9%	12,9%	9,6%	10,1%	13,3%	10,3%	11,1%	14,2%	12,7%
KLOCAL – 25%	7,3 %	9,3 %	7,9 %	8,4%	13,5%	10,4%	10,4%	14%	11,1%	10,7%	13,4%	12,1%
P563 – 25%	7,3 %	8,9 %	8 %	8,1%	13,2%	10%	10,1%	13,6%	11,1%	10,9%	14,14%	12,3%
SNR – 25%	7,2 %	9,1 %	7,9 %	7,7%	12,3%	9,5%	10,2%	13,5%	10,1%	9,5%	13%	11,1%
UBML – 25%	7,1 %	9,1 %	7,7 %	7,6%	12,7%	9,4%	11%	14,5%	10,8%	9,6%	13,5%	11,6%
SLPC – 25%	7,6%	9,1 %	8%	8 %	13,1%	9,9%	11%	13,9%	12,2%	11,4%	14,5%	13,2%
KLPC – 25%	7,3%	9,2%	7,7 %	8,6%	14%	10,7%	11,8%	15,6%	12,8%	11,1%	14,1%	12,6%

Tabla 5.12*: resumen de EERs(%) en el análisis comparativo para los experimentos de todas las condiciones de la base de datos SRE 2008.*

	MODELO		TEST		MODELO		TEST		MODELO		TEST		MODELO		TEST	
	Tlf		Tlf		Mic		Mic		Mic		Tlf		Tlf		Mic	
SISTEMA	GMM	GLDS	SV		GMM	GLDS	SV		GMM	GLDS	SV		GMM	GLDS	SV	
KCEP – 25%	3%	2,8%	5,2%	**3,7%**	7,8%	6,7%	10,5%	**8,3%**	13,5%	11,4%	17,9%	**14,3%**	1,7%	0,8%	2,2%	**1,6%**
KLOCAL – 25%	6,5%	2,8%	3,8%	**4,3%**	1,8%	2%	3,5%	**2,4%**	10,8%	7,1%	11,8%	**9,9%**	5,2%	6,5%	6,4%	**6,0%**
P563 – 25%	6,7%	7,3%	3,3%	**5,7%**	3,9%	4,6%	7,3%	**5,2%**	13,8%	9,5%	11,8%	**11,7%**	2,9%	1,3%	4,8%	**3,0%**
SNR – 25%	8,3%	5%	4%	**5,7%**	9,8%	10,9%	12%	**10,8%**	12,8%	9,9%	20%	**14,2%**	15,1%	9,4%	14,2%	**12,9%**
UBML – 25%	9,2%	5,2%	6%	**6,8%**	10,3%	8,1%	12,4%	**10,2%**	5,8%	3,3%	14,4%	**7,8%**	14,9%	6,1%	10,6%	**10,5%**
SLPC – 25%	3,2%	4,9%	2,4%	**3,5%**	6,2%	5,2%	7,9%	**6,4%**	5,7%	7,4%	2,9%	**5,3%**	-1,1%	-0,8%	-1,6%	**-1,2%**
KLPC – 25%	7%	4,4%	6,8%	**6,0%**	-0,5%	-1,3%	0,1%	**-0,5%**	-0,9%	-3,6%	-1,4%	**-2,0%**	1,7%	1,6%	2,4%	**1,9%**
	6,27%	**4,63%**	**4,50%**		**5,61%**	**5,17%**	**7,67%**		**8,79%**	**6,43%**	**11,06**		**5,77%**	**3,56%**	**5,57%**	

Tabla 5.13*: resumen de mejoras del EER (%) en el análisis comparativo para los experimentos de todas las condiciones de la base de datos SRE 2008.*

En las dos tablas anteriores se resumen los valores de EER (tabla superior) y los valores de mejora tras la exclusión del 25% de los scores con peor calidad (tabla inferior). Basándonos en dichas tablas y las curvas DET, se ha llegado a las siguientes conclusiones:

• **Utilidad de las medidas en habla microfónica (condición *mic-mic*).** Observamos que por lo general las medidas tienen una utilidad mayor que en habla telefónica. Las medidas UBML y SNR siguen manteniendo los valores más altos de mejora del EER, mientras que la P563 disminuye ligeramente.

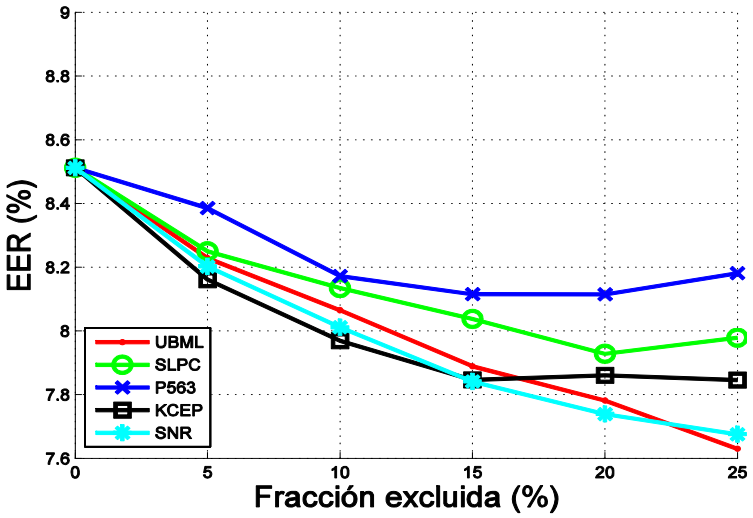
• **Utilidad en cruces (condiciones *mic-tlf y tlf-mic*).** En los cruces micrófono-teléfono son donde las medidas ayudan a mejorar en mayor medida el rendimiento de los sistemas. Observamos que la medida SLPC, que es útil para experimentos de habla microfónica, muestra utilidad en cruces micrófono-teléfono, pero no en teléfono-micrófono. La medida KLPC, que es útil para experimentos de habla telefónica, muestra utilidad en cruces teléfono-micrófono, pero no en micrófono-teléfono.

• **Sistemas**: el sistema GLDS, que tiene la tasa de error más alta en todos los experimentos, es también el que experimenta mejoras más bajas en todos los casos, excepto en experimentos de habla sólo telefónica. Esta diferencia se nota especialmente en los cruces. El sistema SV tiene el mejor comportamiento en habla microfónica al igual que en experimentos de cruce (micrófono-teléfono).

• **Curvas Error vs exclusión**

En las siguientes gráficas se muestra el rendimiento de los sistemas bajo estudio para diferentes fracciones de scores excluidos. Las fracciones de exclusión escogidas han sido las siguientes: 5%, 10%, 15%, 20% y 25%. En este caso se estudian sólo las condiciones que incluyen habla de tipo microfónico, ya que para habla sólo telefónica ya se estudió en la sección 5.1. Todas las curvas corresponden a los resultados obtenidos con el sistema GMM.

MICRÓFONO - MICRÓFONO



MICRÓFONO - TELÉFONO

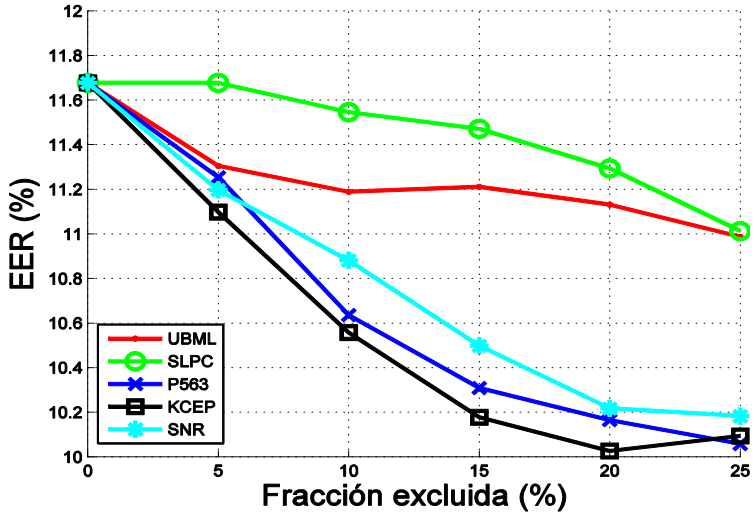


Fig 5.15.a: *curvas Error vs Exclusión del el sistema GMM, para las condiciones micrófono-micrófono y micrófono-teléfono de la base de datos SRE 2008.*

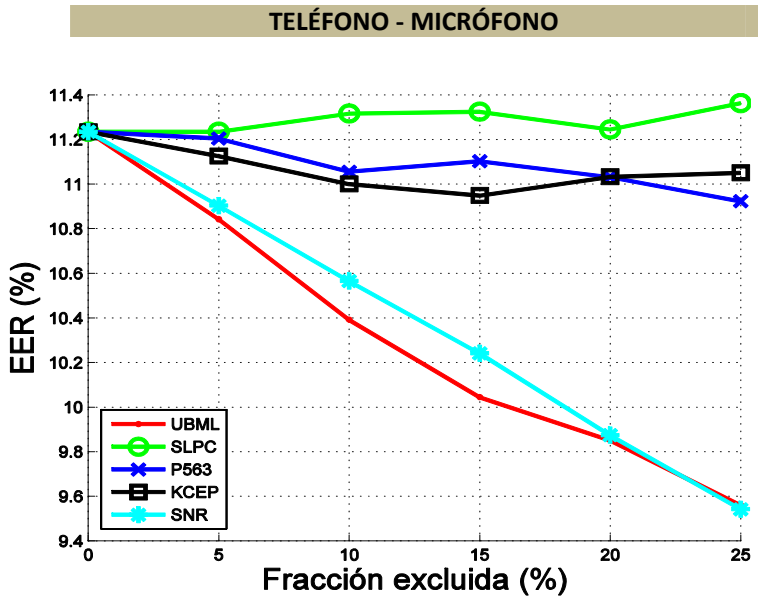


Fig 5.15.b: *curvas Error vs exclusión del sistema GMM para las condiciones micrófono-micrófono y micrófo-teléfono de la base de datos SRE 2008.*

Se observa que el comportamiento es coherente con el análisis realizado con las curvas DET para todas las medidas.

Para aquellas medidas que muestran mejor comportamiento la tendencia de las curvas es uniforme, observándose decrecimientos similares en cada intervalo. Para algunas medidas se observa que el descenso del EER es mayor para las primeras exclusiones de scores. Un claro ejemplo son las medidas de los experimentos de habla sólo microfónica, o las medidas SNR, P563 y KCEP en el cruce micrófono-teléfono. Este efecto ya fue observado en los experimentos de habla telefónica y parece indicar que en un amplio número de casos las medidas que estamos utilizando indican un impacto notablemente superior para los valores más bajos de las mismas. Este efecto podría ser explicado por la distribución que tienen las medidas (ver Anexo A): dado que las calidades bajas coinciden con las colas de la distribución de la población, en las primeras exclusiones se reduce más rápidamente la calidad media de los scores restantes, y como consecuencia el EER disminuye más rápidamente.

5.3 Experimentos con bases de datos forenses

El reconocimiento de locutor con bases de datos forenses presenta una serie de particularidades que hacen más difícil el reconocimiento de personas por la voz. Dichas particularidades tienen que ver la gran mayoría con el hecho de que los factores que influyen en la calidad de la voz están bastante menos controlados que en bases de datos grabadas y preparadas con fines experimentales (distintos dispositivos, el entorno, falta de control sobre el individuo, estado emocional, etc.), lo que resulta en una variabilidad de la calidad de la voz mayor. Por lo tanto, poder medir dicha calidad (para compensarla más adelante) es una tarea difícil que a su vez ofrece un gran potencial en este tipo de habla.

El objetivo de esta sección será experimentar con las medidas de calidad de los experimentos anteriores para conocer la utilidad en una base de datos de este tipo, y extraer toda la información posible que pueda ser de ayuda en futuros experimentos con bases de datos forenses.

La base de datos con la que vamos a trabajar es Ahumada III (sección 4.1). Una de las características de esta base de datos es la forma en la que fue elaborada. Las locuciones están formadas por fragmentos únicamente de voz (se eliminan los silencios), lo cual impide obtener mediciones para aquellos IDs que necesiten los silencios para determinar la calidad de la señal de voz, a saber: SNR, SNR Wiener, P563, Kurtosis global y Skewness global. Por lo tanto los IDs que estudiaremos son los que se listan a continuación:

- Kurtosis local
- Skewness local
- Kurtosis LPC
- Skewness LPC
- Kurtosis Cepstral
- Skewness Cepstral
- UBML

5.3.1 Estudio de indicadores de degradación

Siguiendo la metodología establecida, se obtuvieron las gráficas EER vs Magnitud para los IDs estudiados en esta sección.

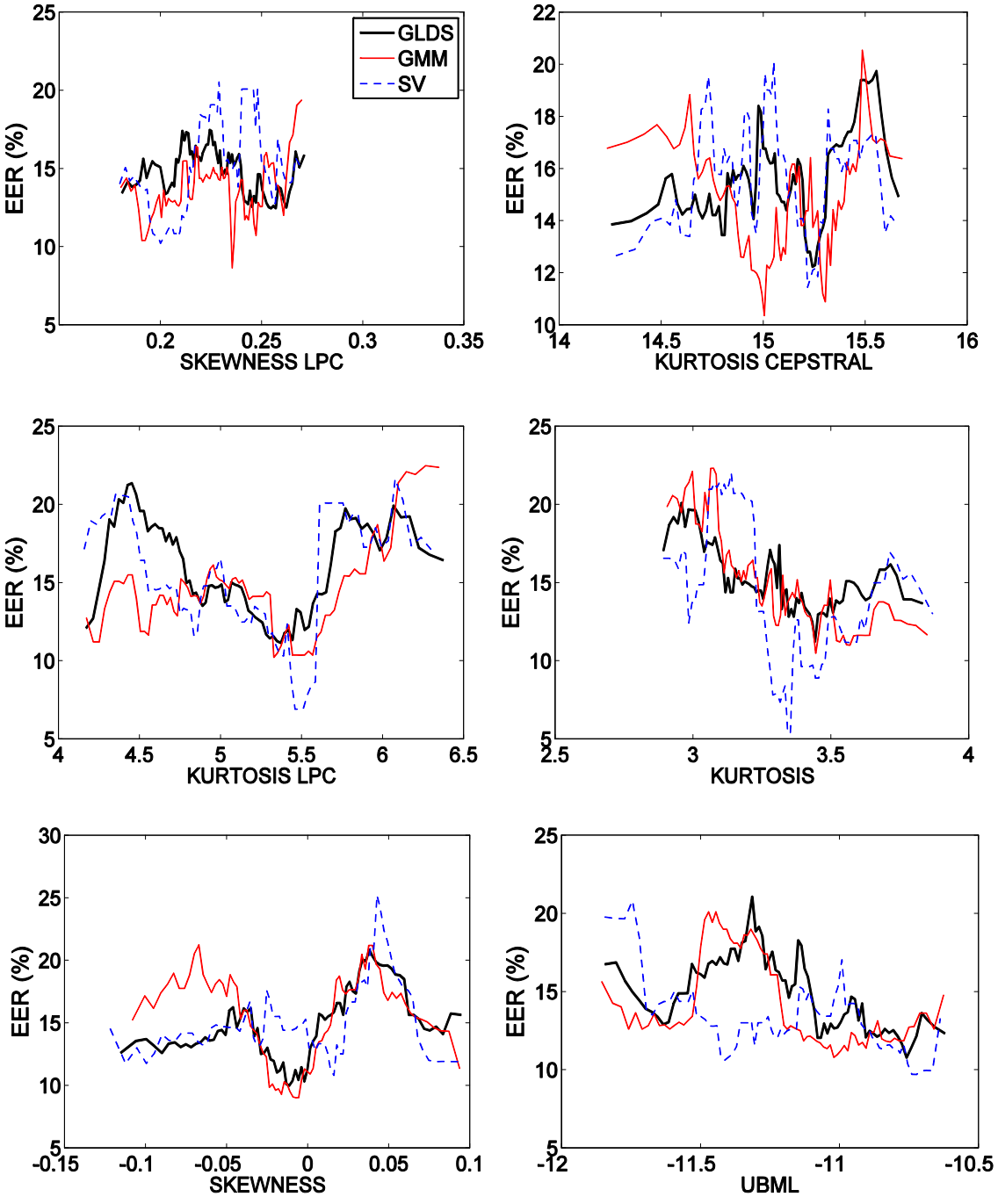


Fig. 5.16.a: *curva Rendimiento vs Magnitud para el indicador Skewness Cepstral. Base de datos Ahumada III.*

.

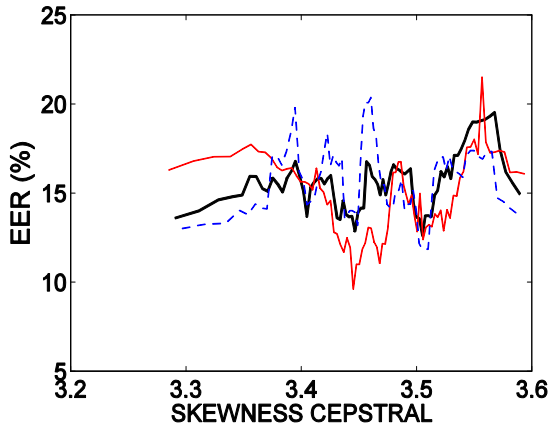


Fig. 5.16.b: *curva Rendimiento vs Magnitud para el indicador Skewness Cepstral.Base de datos Ahumada III.*

Observamos que, por lo general, no existe una relación clara entre los distintos indicadores y el rendimiento de los sistemas. Para algunos de los indicadores (Kurtosis LPC, Kurtosis Local, UBML) se observa una tendencia similar a las que se venían dando para los experimentos anteriores, aunque en ningún caso tan marcada.

Como hemos comentado, el hecho de trabajar con una base de datos forense implica una serie de factores, que podrían explicar este comportamiento de las gráficas:

• **Variabilidad**: las condiciones en las cuales se registraron las distintas locuciones pueden ser muy diversas (teléfono fijo o móvil, espacios abiertos o cerrados, estados emocionales alterados, etc.).

• **Constitución de la base de datos:** las locuciones de testeo tienen una duración de 10 o 15 segundos, con lo que contienen menos información que en el caso de experimentos anteriores. A menor cantidad de información, menor es la probabilidad de obtener una medida de calidad precisa. Por otro lado, los modelos son generados a partir de locuciones de 2 minutos, en condiciones muy diversas de grabación. Esta variabilidad y la gran cantidad de cortes que sufren estos modelos podrían explicar que los valores de los indicadores de degradación no reflejen un valor fiel de la calidad.

• Algunas de las medidas que mostraron un alto impacto en los experimentos anteriores (SNR y P563) no pueden ser calculadas por no disponer de **los silencios de las locuciones**.

5.3.2 Experimentos de correlación

Al igual que en los experimentos anteriores, se han generado las gráficas de dispersión junto con sus coeficientes de correlación lineal.

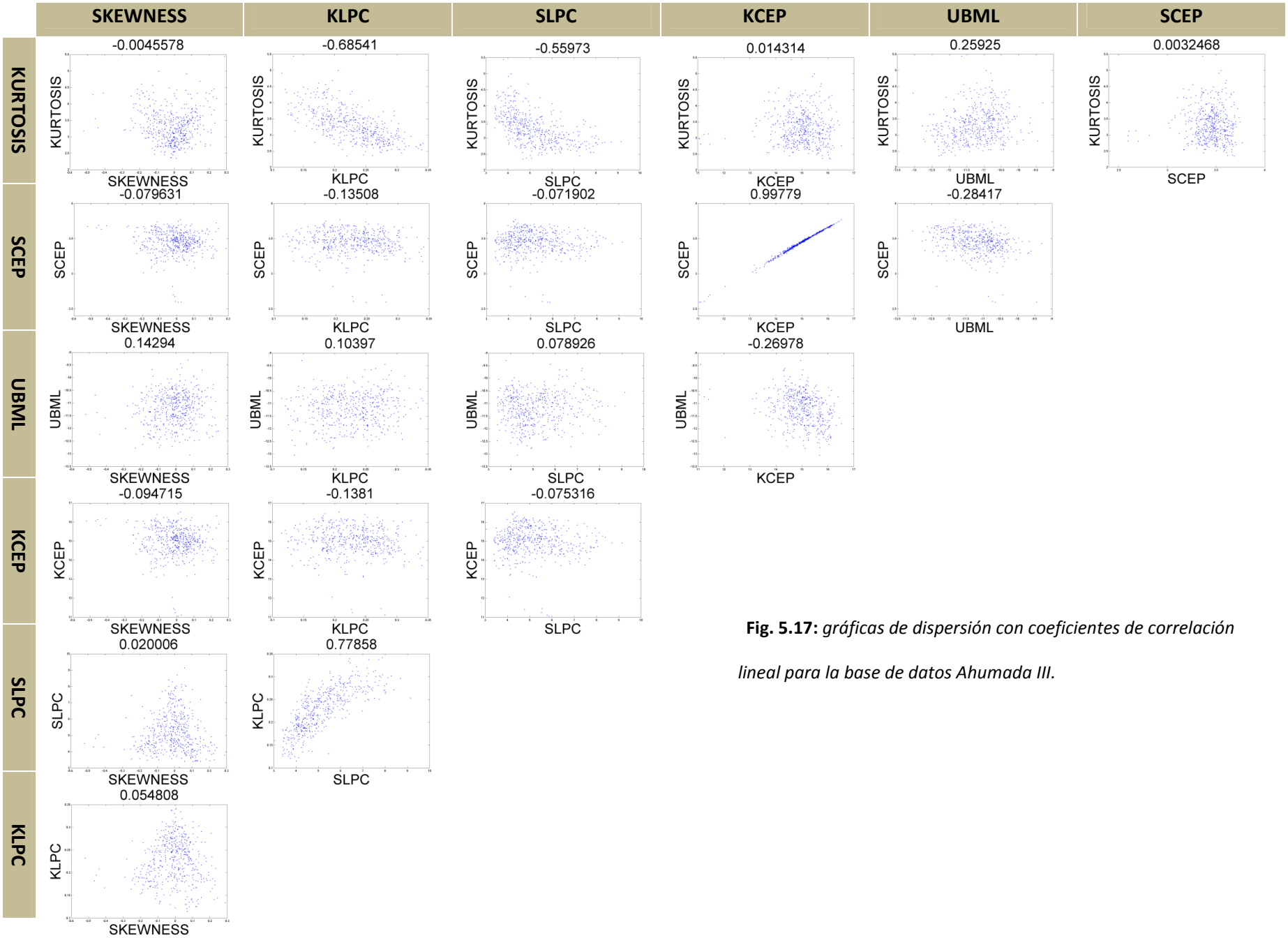


Fig. 5.17: *gráficas de dispersión con coeficientes de correlación lineal para la base de datos Ahumada III.*

5.3.3 Experimentos de utilidad

• Selección de medidas y mapeo

Tal y como se ha venido haciendo en experimentos anteriores, haremos una selección de los IDs que serán mapeados y estudiados como medida de calidad, basados en la información que disponemos de estos IDs hasta ahora. Dado que la base experimental no es tan sólida como en experimentos anteriores (menor cantidad de información con mayor variabilidad), se van a incluir más IDs de lo que se hubiera hecho si nos ciñéramos estrictamente a los criterios seguidos hasta ahora (impacto y coeficientes de correlación sección 4.2).

Los experimentos de correlación han servido para descartar una vez más uno de los IDs de estadísticas de parámetros Cepstrales, debido a la alta correlación con el ID Kurtosis Cepstral. A parte de esto, se han incluido todos los IDs que habían demostrado tener utilidad en los experimentos anteriores, a saber: Kurtosis, Kurtosis LPC, Skewness LPC, Kurtosis Cepstral y UBML.

Para dichos IDs se aplicó el mismo mapeo aplicado en experimentos anteriores, ya que todos experimentan la misma tendencia, y los rangos en los que varían están dentro de los rangos de variación considerados anteriormente.

• Curvas DET

Siguiendo el mismo procedimiento, se obtuvieron las curvas DET original y la correspondiente a la exclusión del 25% de los scores con calidades más bajas. A continuación se muestran las gráficas.

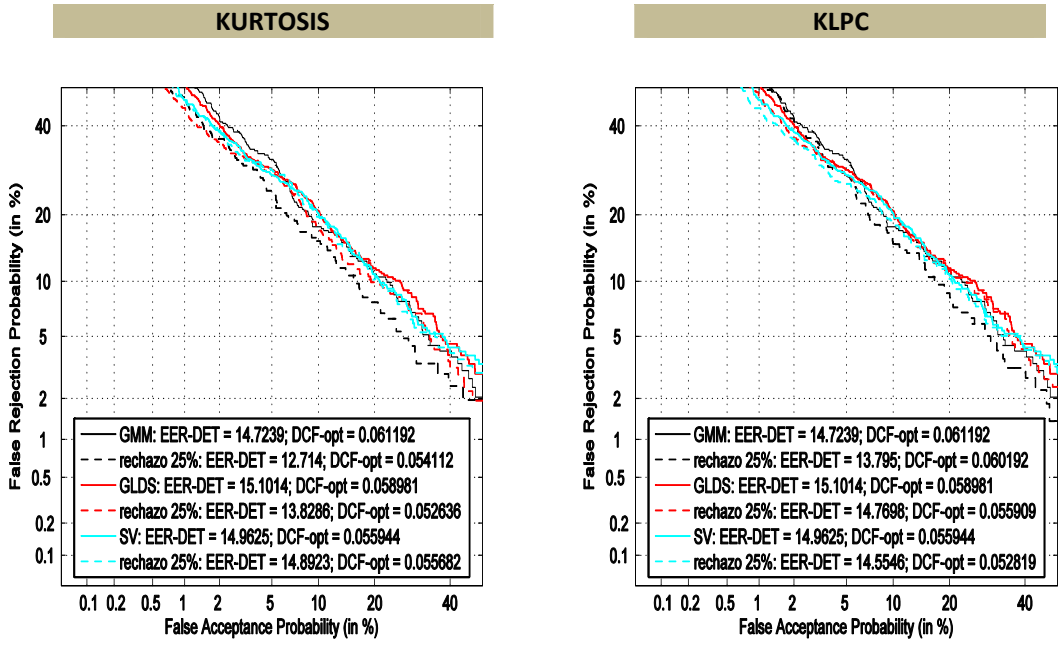


Fig 5.18.a: *curvas DET original (línea continua) y con 25% de scores excluidos (línea discontinua) para los IDs Kurtosis y Kurtosis LPC.*

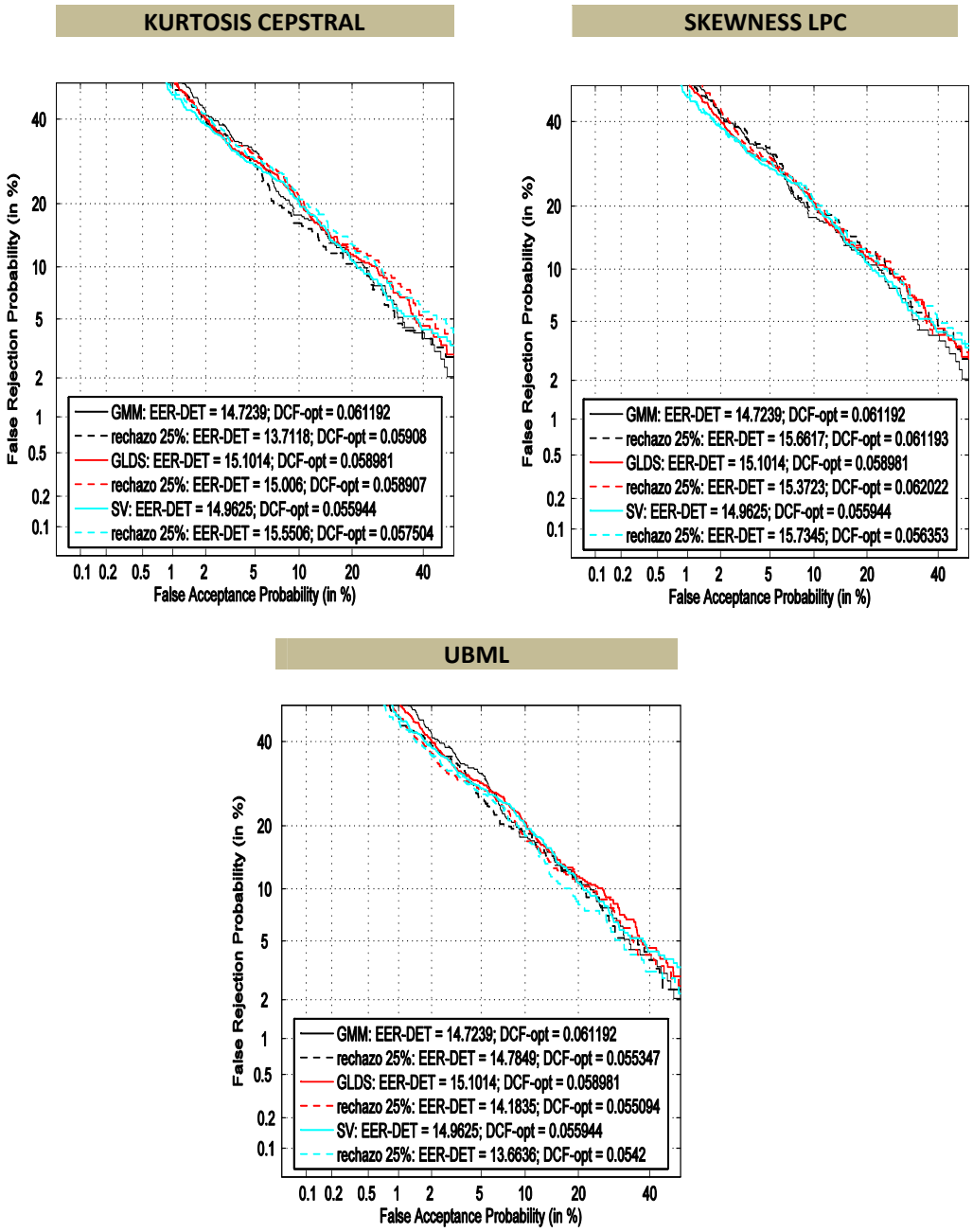


Fig 5.18.b: *curvas DET original (línea continua) y con 25% de scores excluidos (línea discontinua) para los IDs Kurtosis Cepstral, Skewness LPC y UBML.*

Observamos que las mejoras no son tan altas como en experimentos anteriores. Además, al igual que ocurría en las curvas de Rendimiento vs Magnitud, para distintos sistemas bajo una misma medida de calidad, el comportamiento difiere considerablemente. Por ejemplo, para la

medida kurtosis Cepstral, el EER disminuye un punto en el sistema GMM, mientras que para el SVM Super Vector aumenta medio punto.

Para facilitar el análisis, en la siguiente tabla se reúnen los valores de EER(%) para cada una de las curvas:

	SRE 2006		
Sistema / Indicador	GMM	GLDS	SV
SIN RECHAZO	**14,72%**	**15,1%**	**14,96%**
KURTOSIS	12,71%	13,83%	14,89%
KURTOSIS LPC	13,8%	14,77%	14,55%
SKEWNESS LPC	15,66%	15,37%	15,73%
KURTOSIS CEPSTRAL	13,71%	15,00%	15,55%
UBML	14,78%	14,18%	13,66%

Tabla 5.14: *Valores de EER (%) para las curvas DET en la base de datos Ahumada III. Valor original y valor correspondiente a la exclusión del 25% de los scores con menos calidad.*

En la siguiente tabla se muestran las mejoras del EER, en %:

Sistema / Indicador	GMM	GLDS	SV
KURTOSIS	13,65%	8,41%	0,47%
KURTOSIS LPC	6,25%	2,19%	2,74%
SKEWNESS LPC	-6,39%	-1,79%	-5,15%
KURTOSIS CEPSTRAL	6,86%	0,66%	-3,94%
UBML	-0,41%	6,09%	8,69%

Tabla 5.15: *mejoras del EER(%) para las curvas DET*

Por lo general las medidas de calidad que demostraron una utilidad considerable en habla telefónica (Kurtosis, Kurtosis LPC, Kurtosis Cepstral y UBML) siguen teniendo cierta utilidad, mientras que la SLPC, que no demostró tal utilidad, sigue sin ser útil. Por lo tanto parece que en términos de utilidad los resultados son coherentes con los obtenidos en experimentos telefónicos.

Se aprecia que hay un comportamiento dispar entre sistemas para una misma medida de calidad, al igual que entre medidas de calidad para un mismo sistema. Las mismas razones argumentadas para los indicadores de degradación podrían explicar tal disparidad.

• **Curvas "Error vs Exclusión"**

A continuación se muestran las curvas obtenidas, al igual que en los casos anteriores, para un rechazo de 5%, 10%, 15%, 20% y 25% para el sistema GMM.

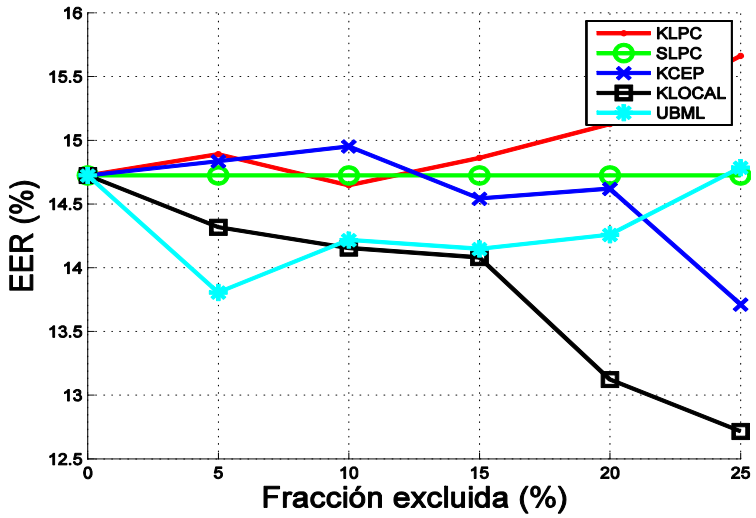


Fig 5.19: *curvas Error vs Exclusión para las medidas de calidad estudiadas en la base de datos Ahumada III.*

Observamos que las curvas son coherentes con las curvas de Magnitud vs Rendimiento, de modo que apenas se aprecia una utilidad muy leve o nula en todas las medidas, siendo la Kurtosis Local y Kurtosis Cepstral las que experimentan mejores resultados.

6 CONCLUSIONES Y TRABAJO FUTURO

6.1 Conclusiones

En este trabajo se han estudiado distintos métodos de estimación de la calidad de la señal de voz en sistemas reconocimiento de locutor.

Con el objetivo de dotar al estudio de una base experimental sólida y extraer la mayor cantidad de información posible, se han utilizado dos bases de datos de común uso en el estado del arte SRE 2006 y SRE 2008) además de una base de datos forense con conversaciones reales de una gran variabilidad en términos de calidad. Además se han estudiado las medidas de calidad en los dos tipos de canal más habituales (teléfono y micrófono). Todas las pruebas han sido realizadas con tres sistemas acústicos diferentes (GMM, SVM-GLDS y SVM-SV), lo que permitió observar la coherencia entre sistemas que utilizan diferente información de la señal de voz, lo cual persigue también reforzar la base experimental del trabajo.

A través del estudio de los indicadores de degradación se aproximó la relación entre estos IDs y el rendimiento de los distintos sistemas, al igual que se conocieron los rangos de variación de los mismos. Todo ello permitió establecer unas funciones de mapeo a medida de calidad a partir de las primeras bases de datos telefónicas estudiadas, y que se mantendrían para el resto de experimentos. De todas las medidas de calidad, han destacado dos por su alto impacto con cualquier base de datos y sistema de los estudiados: SNR y UBML, que a su vez han mostrado una correlación notable entre ellas. La medida propuesta UBML ha mostrado un comportamiento excepcional, siendo la medida más útil en la mayoría de los casos. Dicha medida junto con el estudio de las medidas SNR, P563 y KLPC, fueron presentadas en el International Conference on Biometrics de junio de 2009 en el artículo “Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification”, que se adjunta en el Anexo B.

Los experimentos de correlación nos proporcionaron información que será útil en el futuro a la hora de combinar las medidas de calidad con el fin de compensar la calidad.

Los experimentos de utilidad, además de determinar la utilidad final de las medidas de calidad, han proporcionado datos de interés sobre estas, como por ejemplo el hecho de que las comparaciones target son especialmente sensibles a las variaciones de calidad, o cómo la distribución de la calidad sobre la base de datos bajo estudio influye en los resultados obtenidos.

En experimentos con datos microfónicos y cruce de canal se observó que las medidas más fuertes en experimentos telefónicos (P563, SNR y UBML) seguían mostrando una clara utilidad, mostrando siempre una reducción del EER al excluir scores con peor calidad. Sin embargo se observó que algunas medidas variaban su utilidad en dichos experimentos, tanto para aumentar como para disminuir. Esto permitió conocer medidas de alta utilidad en bases de datos microfónicas (como la SLPC) que no destacaban en habla telefónica. También permitió conocer que el impacto de la calidad es especialmente alto en cruces de modelo microfónicos

y locuciones de testeo telefónicas, donde se llegaron a experimentar mejoras del 20% del EER al excluir el 20% de los scores con peor calidad.

En los experimentos con la base de datos forense Ahumada III no se pudieron utilizar todas las medidas de calidad estudiadas previamente, debido a las condiciones de registro de la base de datos (principalmente por la falta de los silencios). La medida de calidad UBML no mostró la misma utilidad que en los experimentos anteriores, probablemente debido a que el modelo de habla universal había sido entrenado con datos de NIST SRE 2006, que no representaban el entorno experimental para una base de datos forense, pues habían sido registrados en condiciones diferentes (entornos de grabación controlados, reparto de géneros equitativo, etc.) y en un idioma diferente. Otro posible problema con esta base de datos forense es las dimensiones de la misma, pues recoge datos de sólo 61 locutores, y se realiza un número de comparaciones muy inferior a los experimentos anteriores.

A continuación se citan las aportaciones que se han realizado en los distintos puntos que se plantearon en los objetivos del proyecto:

1. **Estudiar el estado del arte**. Se han documentado las últimas tecnologías utilizadas en sistemas de reconocimiento de locutor, algunas de las cuales se utilizan en los sistemas de este trabajo. Se han recopilado los principales métodos de estimación de calidad en SRLs de los escasos estudios que existen en la materia, al igual que otras técnicas de monitorización de la calidad de servicio que pueden ser aplicadas en el caso que nos ocupa.
2. **Implementación de medidas de calidad y proponer nuevas medidas**. Se han implementado un total de 13 medidas de calidad en experimentos telefónicos y 12 en microfónico y cruces. De todos ellos sólo se pudieron probar 7 en bases de datos forenses debido a las limitaciones de la base de datos, lo cual muestra la dificultad para experimentar con este tipo de bases de datos.
   De las 13 medidas estudiadas inicialmente:
   - 5 fueron extraídas de otros estudios de calidad en SRLs, de las cuales sólo dos mostraron ser útiles (P.563 y SNR).
   - 4 fueron extraídas de la documentación de la ITU-T P.563.
   - 4 fueron propuestas en este trabajo, de las cuales una de ellas no mostró utilidad alguna, por lo que fue desechada (SNR Wiener).

3. **Evaluar las medidas de calidad basándonos en protocolos estándar**. El análisis llevado a cabo estuvo basado en el mismo marco teórico de medición que otros estudios del estado del arte [Alonso *et. al*, 2008;Grother *et. al.* 2008] y recomendados por NIST, al igual que se utilizaron métodos de análisis extraídos de este tipo de estudios (experimentos de utilidad y experimentos de correlación) y se ha complementado con el nuevo método de análisis propuesto, el estudio de los indicadores de degradación. Todos estos métodos han permitido realizar un análisis exhaustivo de las medidas bajo diferentes sistemas, bases de datos y protocolos, como ya se ha comentado.

6.2 Líneas de trabajo futuro

Las futuras líneas de trabajo estarían orientadas a utilizar la información obtenida en este PFC para implementar métodos de compensación de calidad en sistemas de reconocimiento de locutor, tales como calibración dependiente de calidad, o cálculo del score dependiente de calidad, con el objetivo de mejorar el rendimiento de los sistemas. La prometedora medida de calidad UBML también será objeto de estudio.

Otra línea de trabajo importante será la obtención de nuevos métodos de estimación de calidad, para lo cual la documentación de la recomendación ITU-P.563 puede ser de ayuda pues recoge hasta 51 parámetros para la estimación de la calidad de la voz.

Como se ha mostrado en este PFC, los experimentos con bases de datos forenses requieren de una base experimental sólida, con suficiente cantidad de información tanto para el entrenamiento de los modelos de habla universal, como para las comparaciones que se realicen. En este campo la calidad juega un papel fundamental, y los datos son difíciles de conseguir, por tanto es una línea de trabajo en la cual se deberían invertir esfuerzos.

7 Referencias

F.Alonso-Fernandez, F. Roli, G. Marcialis, J. Fierrez, J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Performance of fingerprint quality measures depending on sensor technology", SPIE Journal of Electronic Imaging, Special Section on Biometrics: Advances in Security, Usability and Interoperability, Vol. 17, n. 1, January-March 2008

F.Alonso-Fernandez, Biometric Sample Quality and its Application to Multimodal Authentication Systems, Universidad Politecnica de Madrid, October 2008b

M.Alvin / Doddington, George / Kamm, Terri / Ordowski, Mark / Przybocki, Mark (1997): "The DET curve in assessment of detection task performance", In EUROSPEECH-1997, 1895-1898

F.Bimbot,J.-F. Bonastre,C. Fredouille,G. Gravier,I. Magrin-Chagnolleau,Sylvain Meignier,T. Merlin,J. Ortega-Garcıa,Dijana Petrovska-Delacretaz,and Douglas A. Reynolds, A Tutorial on Text-Independent Speaker Verification, 2004.

C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

W.M.Campbell, D. E. Sturim y D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters, IEEE, Vol. 13, N. 5, pp. 308-311, Mayo 2006

J.P. Campbell, H. Nakasone, C. Cieri, D. Miller,K. Walker, A. F. Martin, and M. A. Przybocki, “The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation,” in Proc. of Odyssey, 2004, pp. 29–32.

W.M.Campbell. Generalized linear discriminate sequence kernels for speaker recognition. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 161-164, 2002.

J.Chen, J. Benesty, Y.Huang, E.J. Diethorn, Fundamentals of Noise Reduction, Springer Handbook of Speech Processing, Chapter 43

J.R.Deller, J. H. L. Hansen y J. G. Proakis. Discrete-time processing of speech signals. Wiley-IEEE Press, Septiembre 1999.

L. Ferrer, Martin Graciarena, Argyris Zymnis, Elizabeth Shriberg,, System Combination using Auxiliary Information for Speaker Verification, 2008

J.Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures", Pattern Recognition, Vol. 38, n. 5, pp. 777-779, May 2005

S.Furui. Cepstral Analysis technique for automatic speaker verification. IEEE Transactions on acoustics, speech and signal processing, Vol. ASSP-29, N. 2, Abril 1981.

D.Garcia-Romero, Julian Fierrez-Aguilar,Joaquin Gonzalez-Rodriguez , Javier Ortega-Garcia, Using Quality Measures for Multilevel Speaker Recognition, 2005

A.A.Garcia y R. J. Mammone. Channel-robust speaker identification using modifiedmean cepstral mean normalization with frecuency warping. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 325-328, 1999.

J.Gonzalez-Rodriguez, D.T.Toledano, and J.Ortega-Garcia. Voice Biometrics. Springer 2007.

V.Grancharov, W.B. Hleijn, Speech Quality Assessment, Springer Handbook of Speech Processing, Chapter 5, 2007

P.Gray, M.P. Hollier y R.E. Massara, Non-intrusive speech quality assessment using vocal-tract models, 2000

P.Grother, P., Tabassi, E.: Performance of Biometric Quality Measures. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29(4), pp. 531–543 (2007)

H.Hermansky y N. Morgan. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing 2, pp. 578-589, Octubre 1994.

W.Hess: Pitch Determination of Speech Signals (Springer,Berlin, Heidelberg), 1983

Hicklin and R.Khanna, The Role of Data Quality in Biometric Systems, 2006

ITU-P563: http://www.itu.int/itudoc/itu-t/aap/sg12aap/history/p563/index.html

ITU-T Rec. P.800: Methods for subjective determination of transmission quality, 1996.

P.Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification,” in Proc. of ICASSP, 2004, vol. 1, pp. 37–40.

M1/05-0306, Biometric Sample Quality Standard Draft (Revision 4), 6.

L.Malfait, Jens Berger y Martin Kasner, P.563 – The ITU-T Standard for Single-Ended Speech Quality Assessment, IEEE Transactions on Audio and Language Processing, Vol14, NO. 6,2004

A.Maltoni, D. Maio, A. K. Jain y S. Prabhakar. Handbook of Fingerprint Recognition, Springer 2003.

I.Mateos-García, Máquinas de Vectores Soporte (SVM) para Reconocimiento de Locutor e Idioma, Universidad Autonoma de Madrid, Julio 2007

NIST QUALITY WORKSHOP: http://www.itl.nist.gov/iad/894.03/quality/workshop/

NIST SRE 2006: National Institute of Standards and Technology Speaker Recognition Evaluation: http://www.nist.gov/speech/tests/sre/2006/index.html

NIST SRE 2008: National Institute of Standards and Technology Speaker Recognition Evaluation: http://www.nist.gov/speech/tests/sre/2008/index.html

NIST SRE: Speaker Recognition: http://www.nist.gov/speech/tests/sre/

J.Pelecanos y S. Sridharan. Feature warping for robust speaker verification. Proc. IEEE Speaker and Language Recognition Workshop (Odyssey), pp. 213-218, 2001.

D.Ramos, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Universidad Autonoma de Madrid, November 2007

D.Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-case database in Spanish", in Proceedings of Interspeech 2008, pp. 1493-1496, September 2008

D.Ramos, J. Gonzalez-Dominguez, D. T. Toledano and J. Gonzalez-Rodriguez, "Speaker Feature", Stan Z. Li (Eds.), Encyclopedia of Biometrics (ISBN 978-0-387-73003-5), Springer, July 2009.

D. Reynolds et al. Supersid project: Exploiting high-level information for high accuracy speaker recognition. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, pp 784-787, Abril 2003.

D.Reynolds, A Tutorial on Text-Independent Speaker Verification, EURASIP Journal on Advances in Signal Processing, 2004.

D.Reynolds, An overview of speaker recognition technology. In Proc. Of ICASSP, pager 4072-4075, 2003.

D.Reynolds. Channel robust speaker verification via feature mapping. Proc. IEEE, Vol. 2, pp. 53-56, International Conference on Acoustics, Speech and Signal Processing (ICASSP),Abril 2003b.

J.Richiardi y Andrzej Drygajlo, Evaluation of speech quality measures for the purpose of speaker verification, 2007

Won Park ,Linear Predictive Speech Processing, Texas A&M University-Kings.

GLOSARIO

GMM	**Gaussian Mixture Model**
ID	**Indicador de degradación**
KBINS	**Kurtosis aproximada por historial en el dominio del tiempo**
KCEP	**Kurtosis sobre coeficientes Cepstrales**
KGLOBAL	**Kurtosis en dominio temporal sobre la locución entera**
KLOCAL	**Kurtosis en dominio temporal sobre tramas pequeñas**
KLPC	**Kurtosis sobre coeficientes LPC**
LPC	**Linear Predictive Coding**
MFCC	**Mel-Frequency Cepstral Coefficients**
SCEP	**Skewness sobre coeficientes Cepstrales**
SGLOBAL	**Skewness en dominio temporal sobre la locución entera**
SLOCAL	**Skewness en dominio temporal sobre tramas pequeñas**
SLPC	**Skewness sobre coeficientes LPC**
SNR	**Relación señal a ruido**
SRL	**Sistema de reconocimiento de locutor**
SVM	**Support Vector Machine**
SVM-GLDS	**Generalized Linear Discriminant Sequence**
SVM-SV	**SVM- Supervector**
UBM	**Universal Background Model (modelo de habla universal)**
UBML	**Universal Background Model Likelihood**

ANEXO A: ESTADÍSTICAS DE LOS INDICADORES DE DEGRADACIÓN.

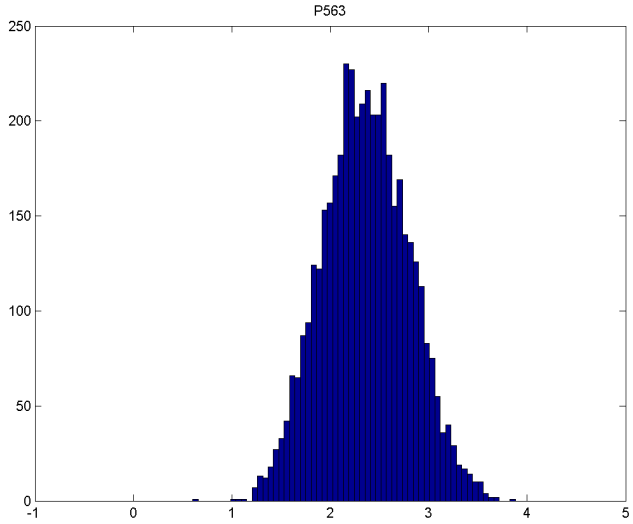
Estadísticas en habla telefónica (SRE 2008).

	Media	Varianza	Media Modelos	Varianza Modelos	Media Tests	Varianza Tests
p563	2,31	0,20	2,32	0,19	2,30	0,21
snr	34,04	65,20	31,83	80,75	35,58	48,64
snr wiener	15,83	5,97	15,81	6,10	15,84	5,88
Kurtosis local	3,89	0,62	3,84	0,87	3,92	0,45
Skewness local	-0,01	0,03	-0,01	0,02	0,00	0,03
klpc	6,05	0,87	6,10	0,81	6,02	0,90
slpc	-0,21	0,07	-0,23	0,07	-0,19	0,08
kcep	13,00	1,17	12,98	1,13	13,02	1,20
scep	3,04	0,05	3,03	0,05	3,04	0,05
kurtosis global	50,09	3657,16	48,02	3254,93	51,53	3933,00
skewness global	-0,03	0,57	-0,05	0,51	-0,02	0,61
kbins	6,25	10,88	6,22	10,77	6,27	10,97
ubml	-9,32	0,48	-9,27	0,43	-9,35	0,51

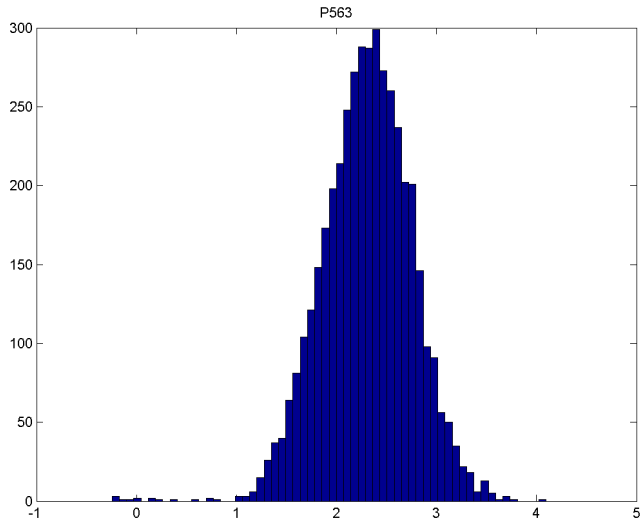
Estadísticas en habla microfónica (SRE 2008).

	Media	Varianza	Media modelos	Varianza Modelos	Media Tests	Varianza Tests
p563	1,89	0,29	1,91	0,39	1,88	0,23
snr	18,88	20,76	18,47	32,94	19,14	12,87
Kurtosis local	2,99	0,16	3,03	0,15	2,96	0,16
Skewness local	0,09	0,02	0,12	0,03	0,07	0,02
klpc	6,69	1,41	6,42	1,30	6,86	1,41
slpc	0,59	0,08	-0,53	0,07	-0,63	0,07
kcep	10,65	6,93	10,12	12,37	10,99	3,19
scep	2,46	0,49	2,28	0,93	2,57	0,18
kurtosis global	29,77	573,44	29,97	583,12	29,64	567,51
skewness global	0,33	1,03	0,34	1,05	0,33	1,01
kbins	2,38	1,43	2,43	1,20	2,35	1,57
ubml	-10,51	0,42	-10,53	0,40	-10,49	0,43

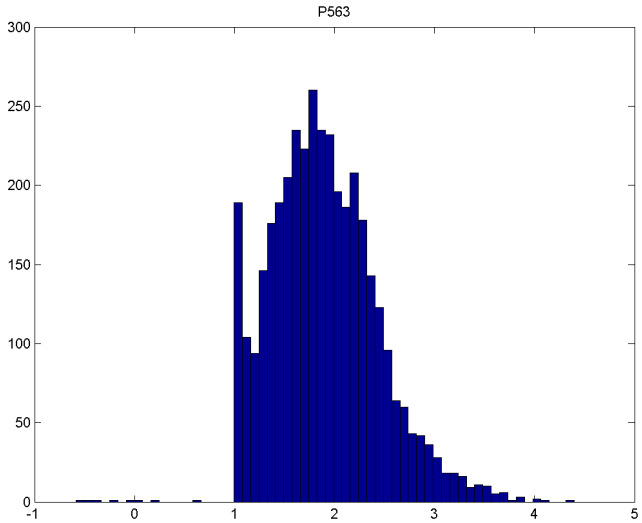
SRE 2006 TELEFÓNICO



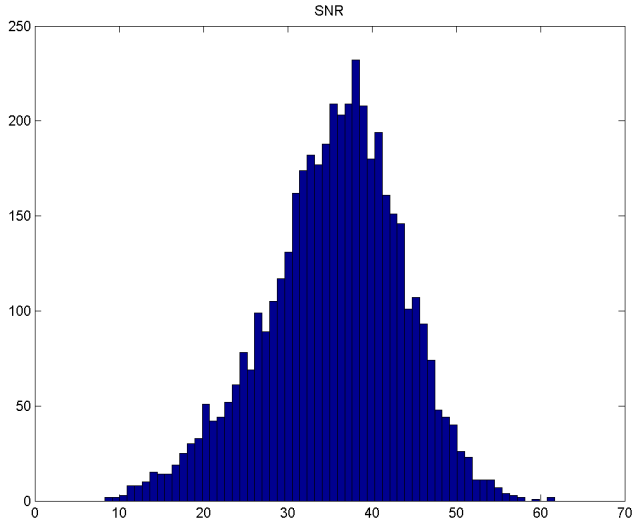
SRE 2008 TELEFÓNICO



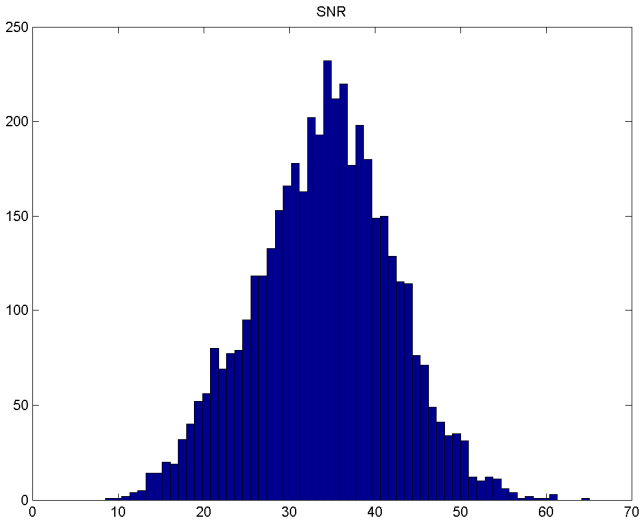
SRE 2008 MICROFÓNICO



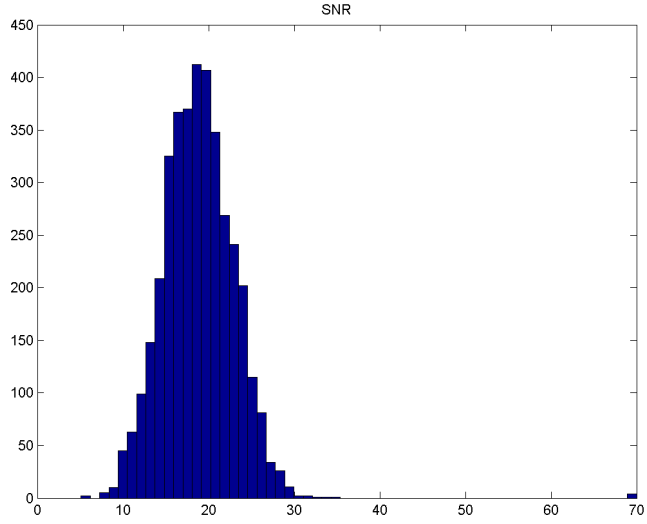
SRE 2006 TELEFÓNICO



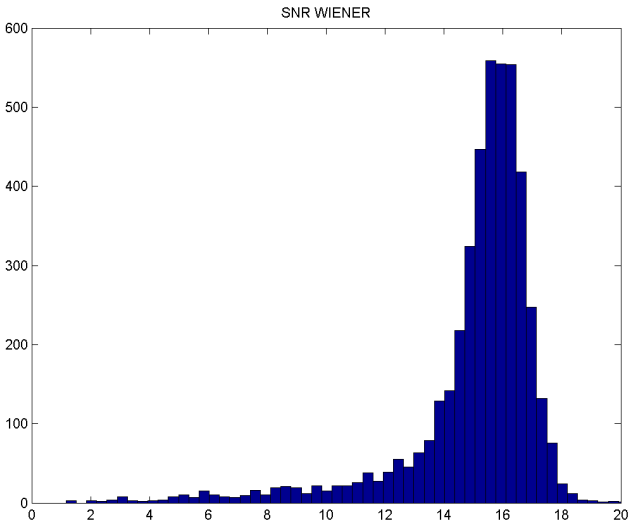
SRE 2008 TELEFÓNICO



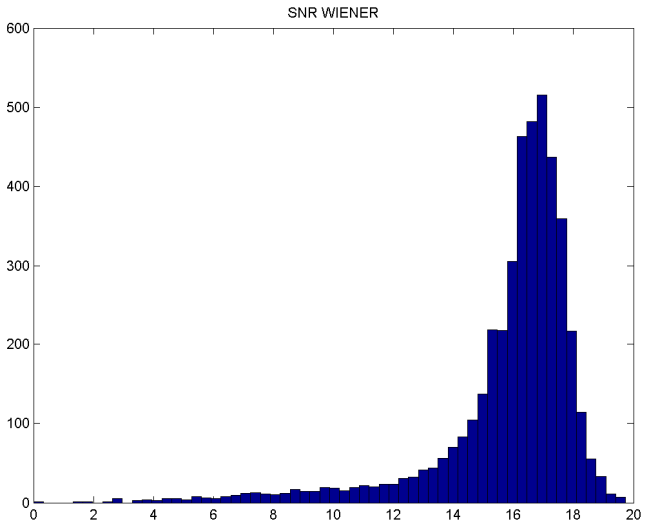
SRE 2008 MICROFÓNICO



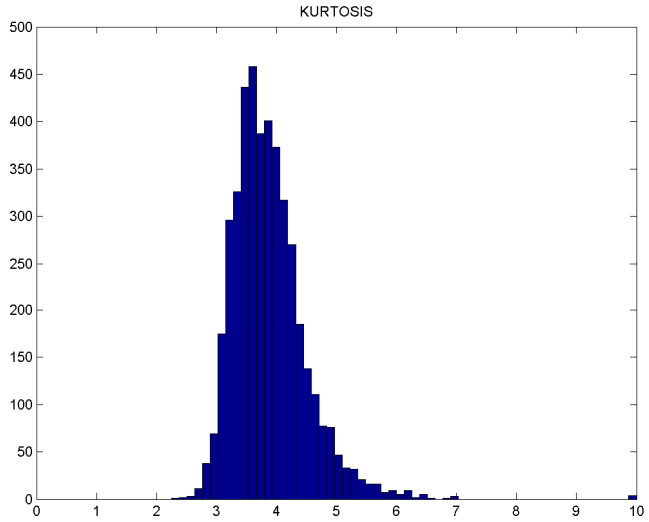
SRE 2006 TELEFÓNICO



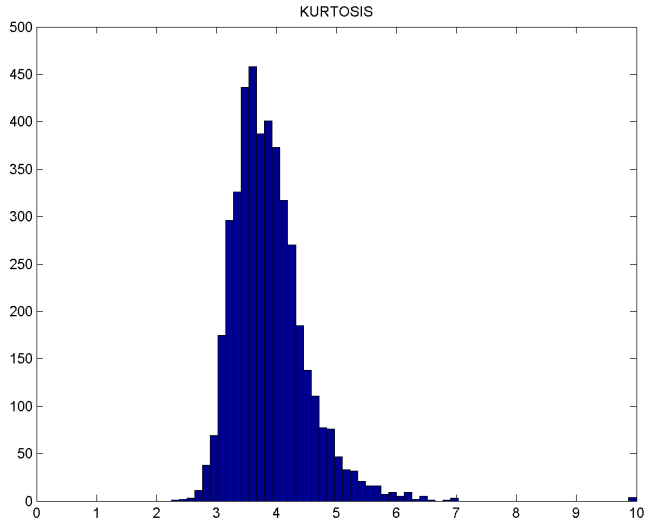
SRE 2006 TELEFÓNICO



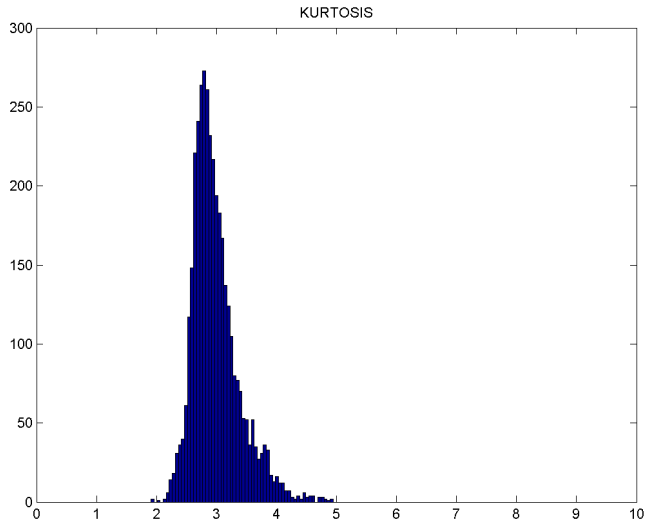
SRE 2006 TELEFÓNICO



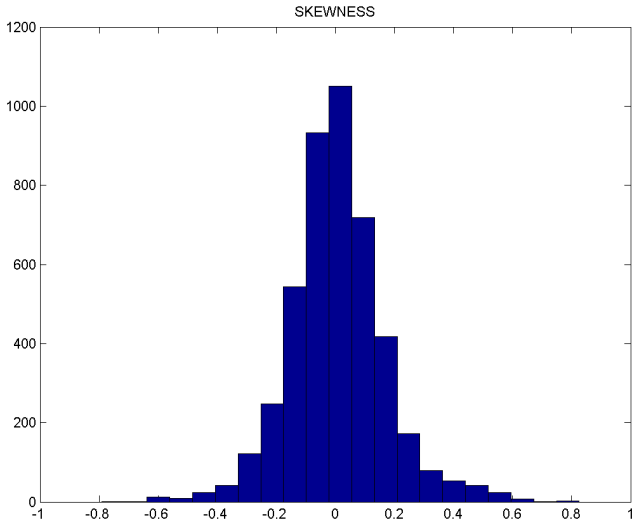
SRE 2006 TELEFÓNICO



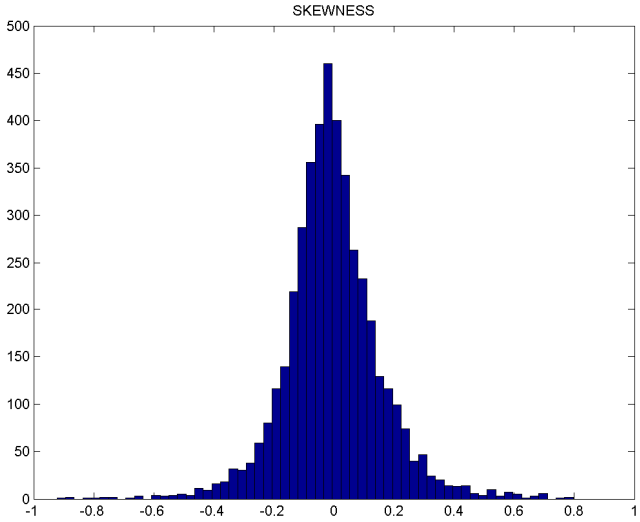
SRE 2006 TELEFÓNICO



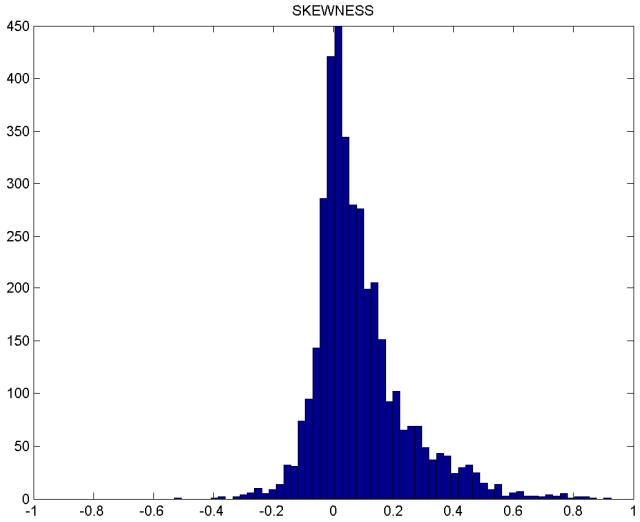
SRE 2006 TELEFÓNICO



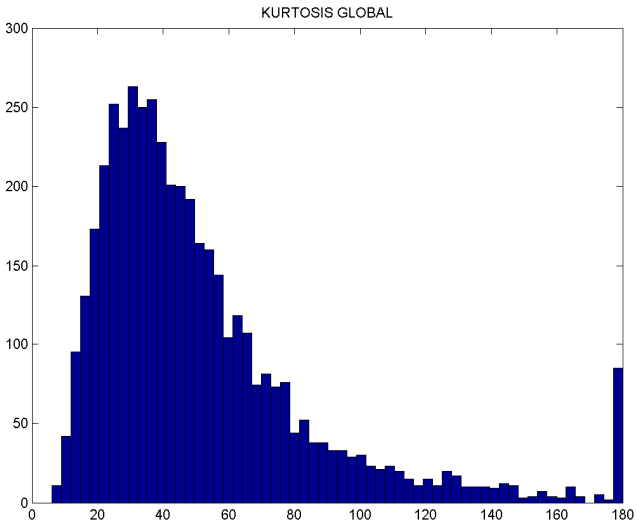
SRE 2006 TELEFÓNICO



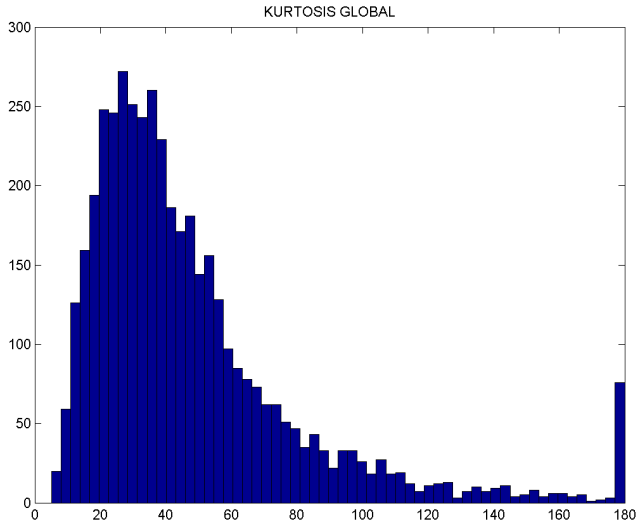
SRE 2006 TELEFÓNICO



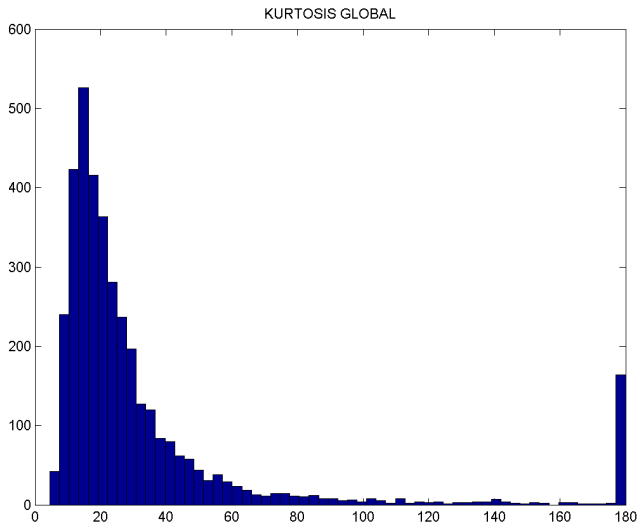
SRE 2006 TELEFÓNICO



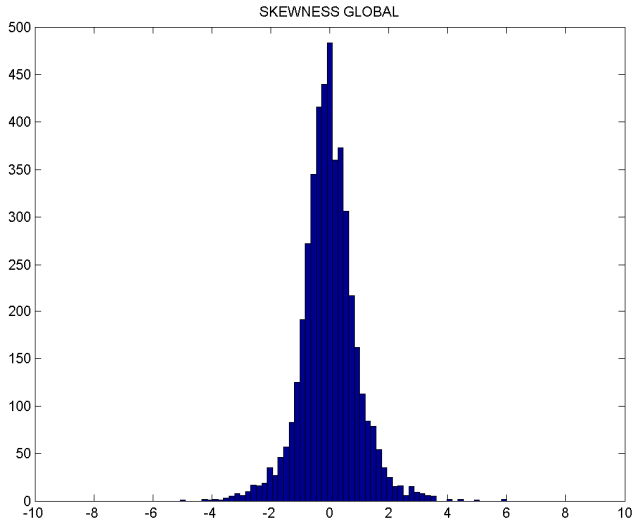
SRE 2006 TELEFÓNICO



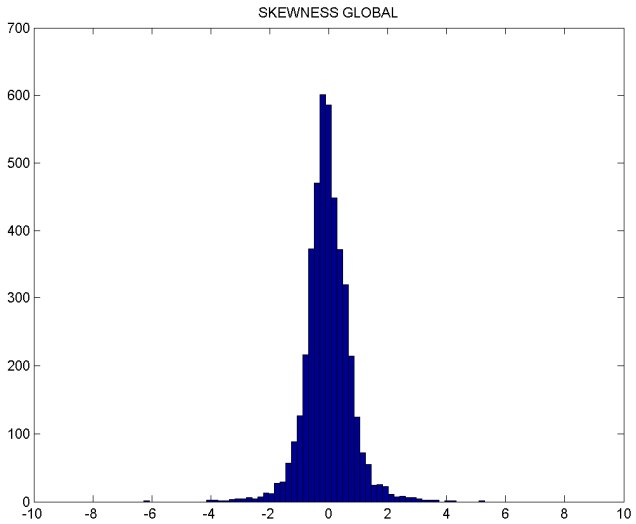
SRE 2006 TELEFÓNICO



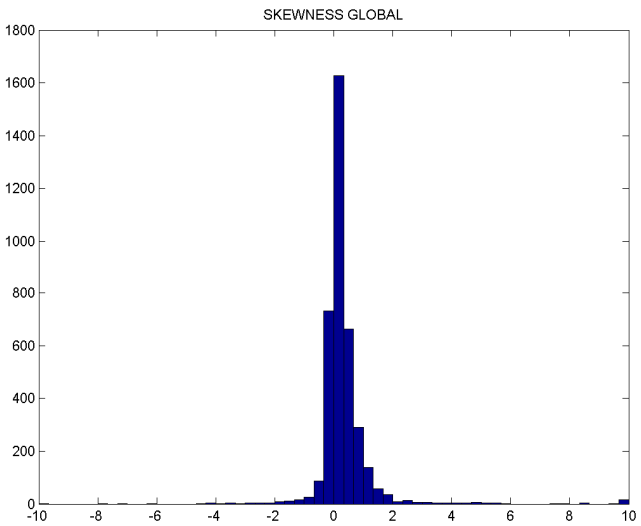
SRE 2006 TELEFÓNICO



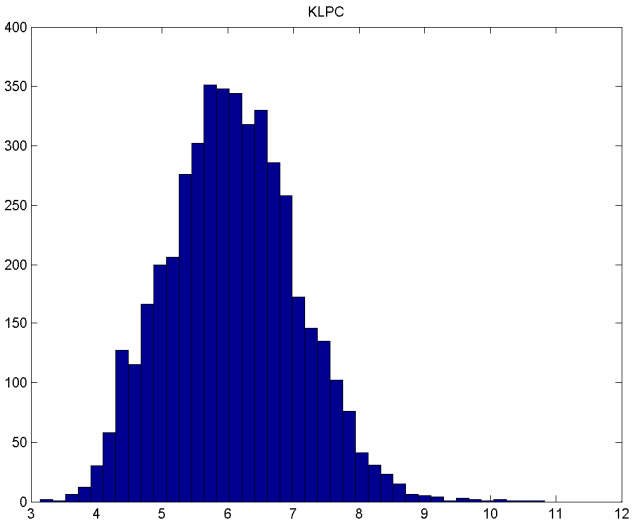
SRE 2006 TELEFÓNICO



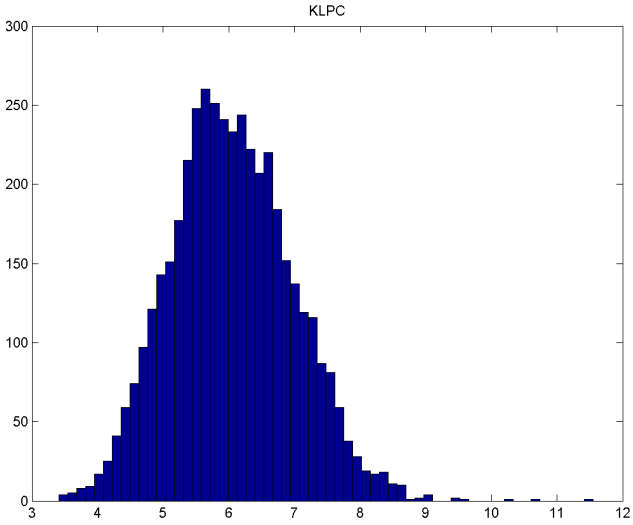
SRE 2006 TELEFÓNICO



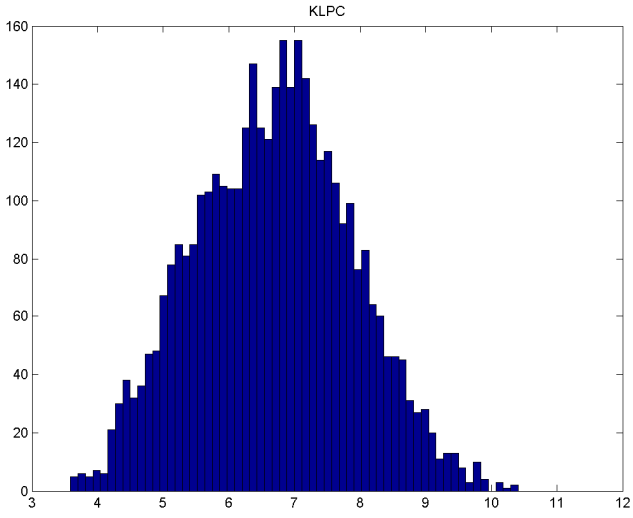
SRE 2006 TELEFÓNICO



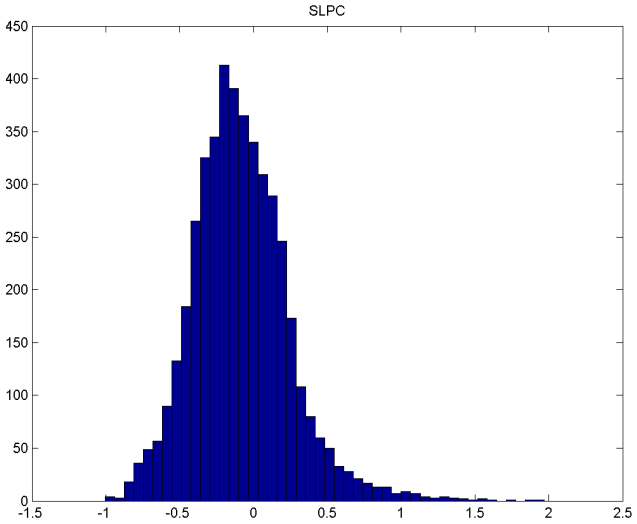
SRE 2006 TELEFÓNICO



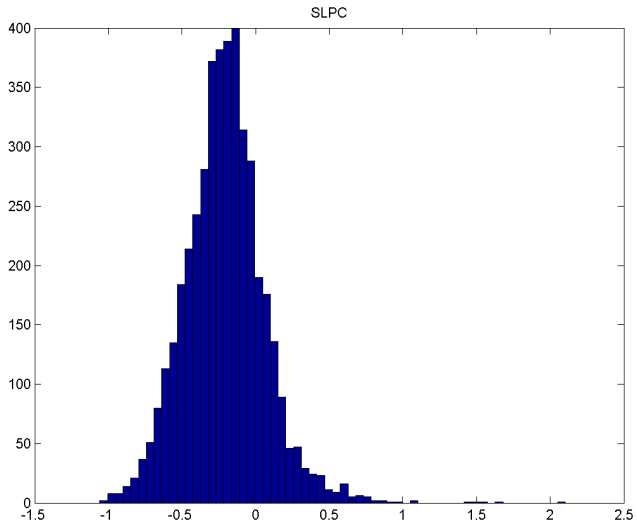
SRE 2006 TELEFÓNICO



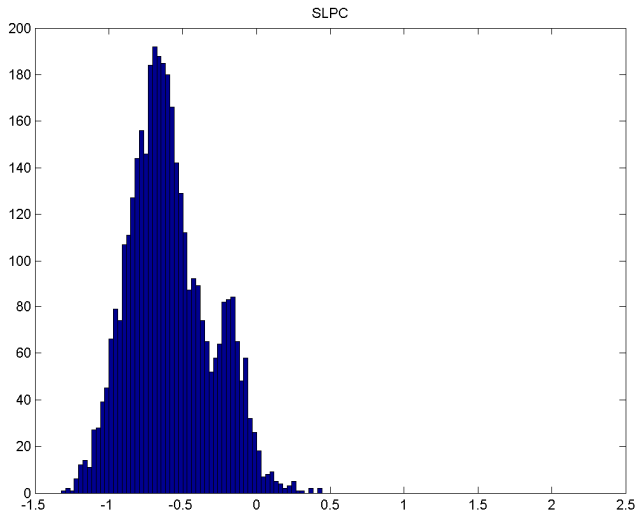
SRE 2006 TELEFÓNICO



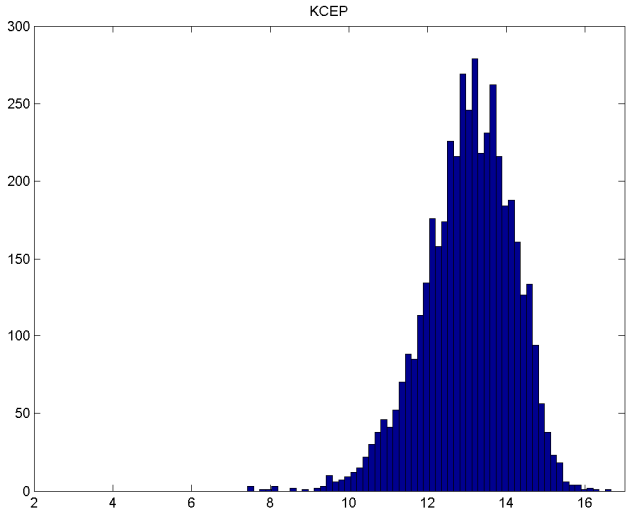
SRE 2006 TELEFÓNICO



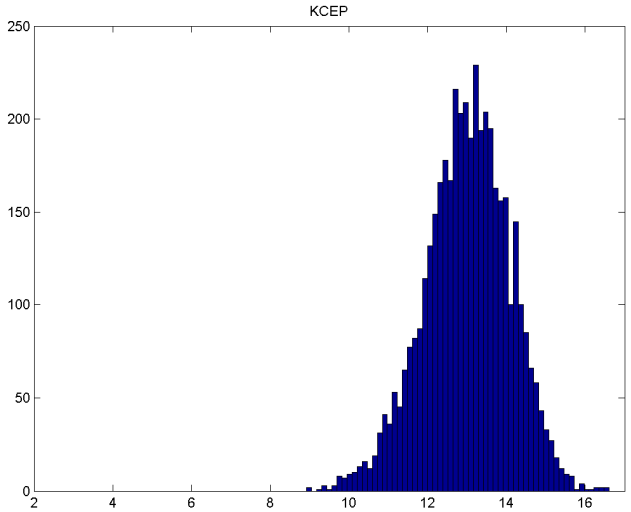
SRE 2006 TELEFÓNICO



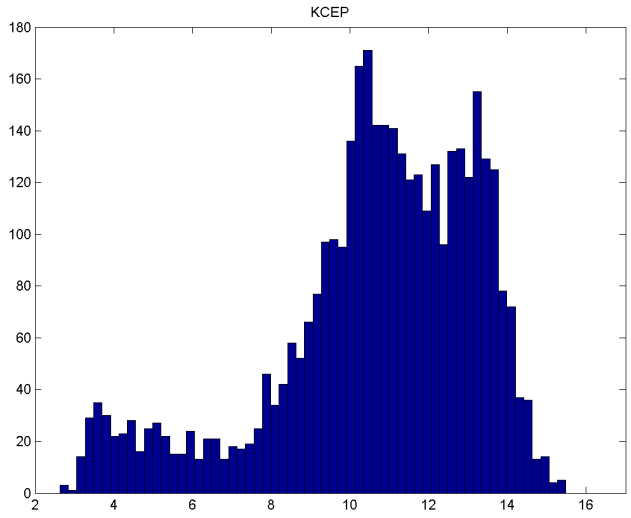
SRE 2006 TELEFÓNICO



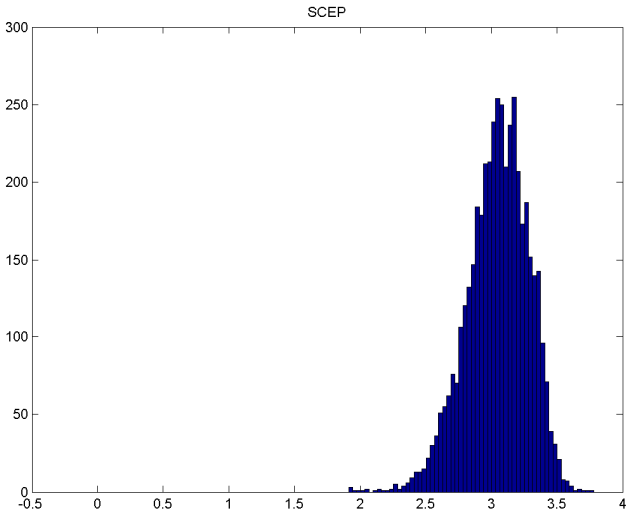
SRE 2006 TELEFÓNICO



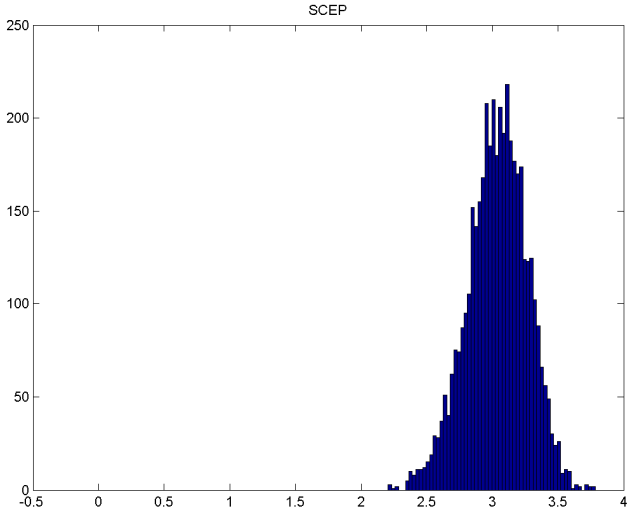
SRE 2006 TELEFÓNICO



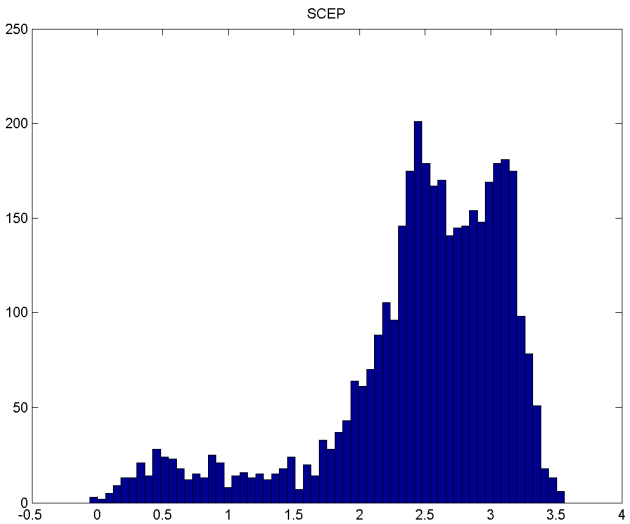
SRE 2006 TELEFÓNICO



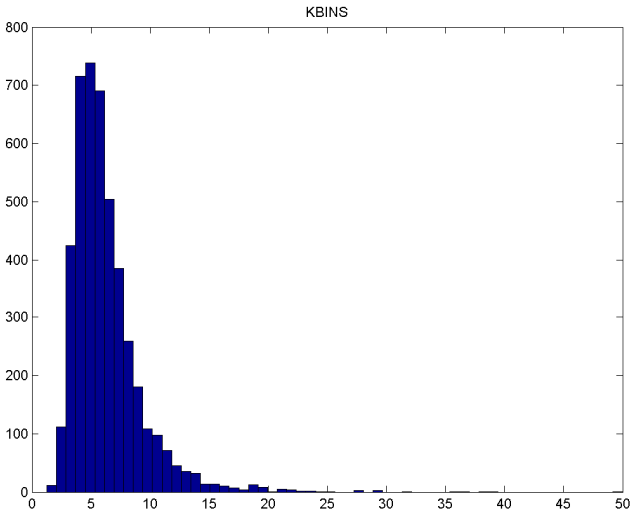
SRE 2006 TELEFÓNICO



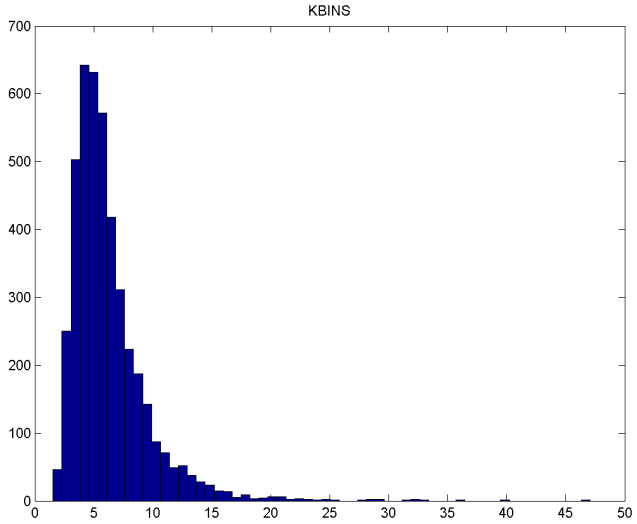
SRE 2006 TELEFÓNICO



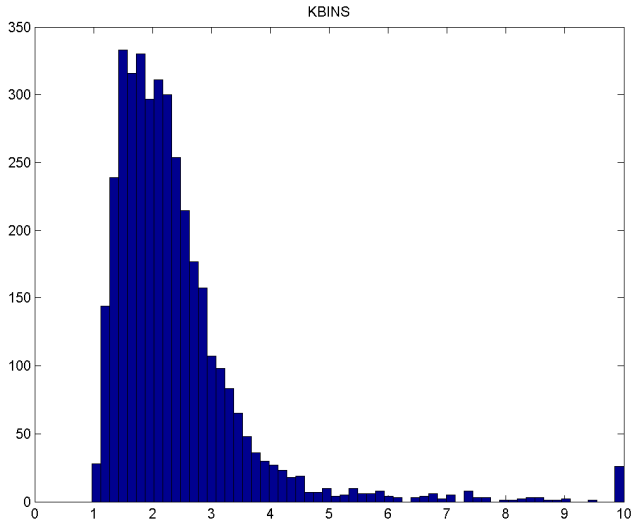
SRE 2006 TELEFÓNICO



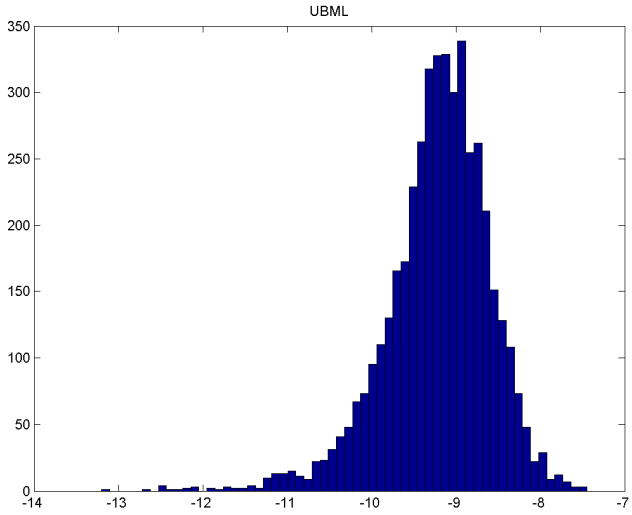
SRE 2006 TELEFÓNICO



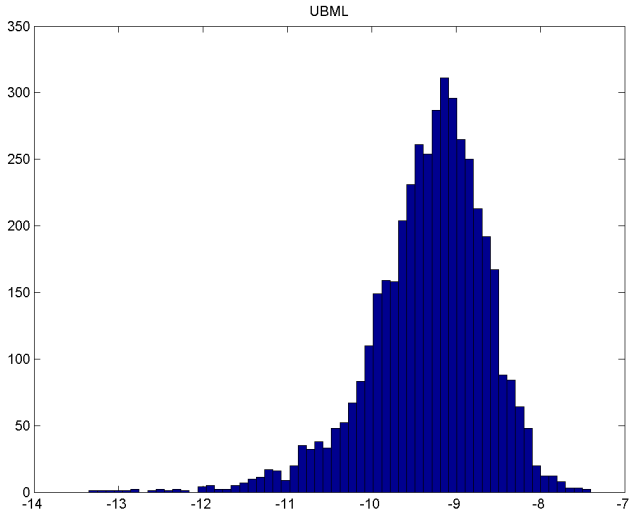
SRE 2006 TELEFÓNICO



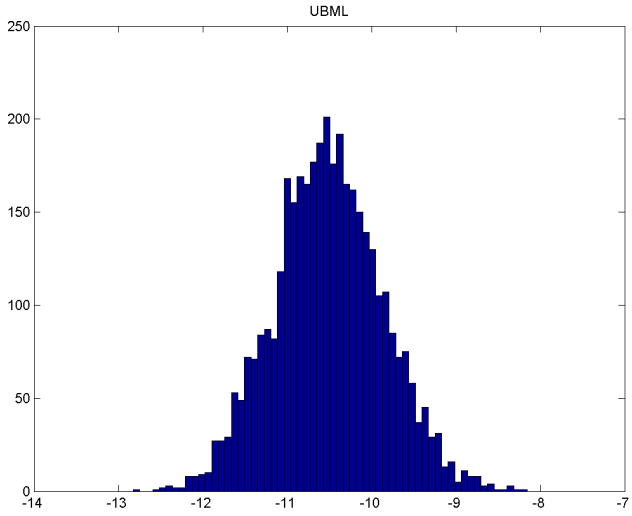
SRE 2006 TELEFÓNICO



SRE 2006 TELEFÓNICO



SRE 2006 TELEFÓNICO



Anexo B: Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification

Alberto Harriero, Daniel Ramos, Joaquin Gonzalez-Rodriguez and Julian Fierrez

ATVS – Biometric Recognition Group.
Escuela Politecnica Superior, Universidad Autonoma de Madrid.
C. Francisco Tomás y Valiente 11, 28049 Madrid, Spain.
{alberto.harriero, daniel.ramos, joaquin.gonzalez, julian.fierrez}@uam.es

Abstract. In this work, we analyze several quality measures for speaker verification from the point of view of their utility, i.e., their ability to predict performance in an authentication task. We select several quality measures derived from classic indicators of speech degradation, namely ITU P.563 estimator of subjective quality, signal to noise ratio and kurtosis of linear predictive coefficients. Moreover, we propose a novel quality measure derived from what we have called Universal Background Model Likelihood (UBML), which indicates the degradation of a speech utterance in terms of its divergence with respect to a given universal model. Utility of quality measures is evaluated following the protocols and databases of NIST Speaker Recognition Evaluation (SRE) 2006 and 2008 (telephone-only subset), and ultimately by means of error-vs.-rejection plots as recommended by NIST. Results presented in this study show significant utility for all the quality measures analyzed, and also a moderate decorrelation among them.

Keywords: speaker verification, quality, utility, SNR, degradation indicator.

1 Introduction

Speaker recognition is nowadays a mature field with multiple applications in security, access control, intelligence and forensics. The State of the Art is based on the use of spectral information of the speech signal, combining such information in multiple ways, and compensating the inter-session variability of speech recordings [1,2].

Despite the significant advance on the performance of the technology in the field, partly due to the efforts of NIST and their successful periodic Speaker Recognition Evaluations [3], the field of speaker recognition faces important challenges. Among them, performance of comparisons when there is a high mismatch between enrollment and testing speech conditions is far from being solved, although the improvements in this sense in the last years have been remarkable [1]. Moreover, the mismatch in the conditions of the speech databases for system tuning and for operational work (the so-called database mismatch problem [4]) has a strong impact in the performance of the systems, and attenuates the beneficial effects of compensation techniques.

In order to solve the problems associated to session variability in speech, the speaker recognition scientific community continues their efforts on improving the existing compensation algorithms [1]. These methods are mainly based on data-driven approaches modeled with statistical techniques such as factor analysis [1]. Although their demonstrated success, such techniques are sensitive to the existence of a rich development corpus, desirably in similar conditions to those of the operational framework, which may not be available in general. Moreover, there

[1] The last research workshop on the topic at John Hoskins University deserves special attention (http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/).

is other knowledge about the speech signal which can be efficiently extracted from excerpts and used as information about the variability of the speech signal and its impact on the performance of speaker recognition systems. Among such knowledge are the quality measures, as recently proposed by NIST [5].

In this work, we present an analysis of several quality measures from the point of view of their utility, i.e., their usefulness as a predictor of system performance. Some of the analyzed quality measures are derived from classical indicators of speech degradation, namely Signal to Noise Ratio (SNR), statistics from Linear Predictive Coefficients (LPC) and estimators of subjective quality (such as ITU P.563 recommendation [6]). Moreover, we propose a quality measure with an attractive interpretation, derived from what we have called Universal Background Model Likelihood (UBML). The work also presents a framework for the obtaining of the proposed quality measures from speech. The paper is completed with experimental results using telephone speech and protocols from recent NIST Speaker Recognition Evaluation Evaluations (SRE), where the utility of quality measures is shown by the performance measures recommended by NIST [5].

The paper is organized as follows. In section 2, we define the quality measurement framework according to previous work in the literature [6,7], We also present three classical quality measures derived from classical indicators of speech degradation. In section 3, we present a novel quality measure based on what we have called the Universal Background Model Likelihood (UBML). Results showing the analysis of the four analyzed quality measures, including the proposed one derived from UBML, are described in section 4, where the utility of the proposed measures is analyzed using two different databases from NIST Speaker Recognition Evaluations (2006 and 2008). Experiments allow the identification of the most useful quality measures for predicting performance, based on protocols recommended by NIST [5]. Finally, conclusions are drawn in section 5.

2 Quality measures for speaker verification

The idea that the quality of the speech signal affects the ability of an automatic system to distinguish among people from their voices is somewhat intuitive, as it happens in other biometric traits [8]. In fact, the measurement of speech quality has been a major topic of research during the last decades [9]. The need to monitor the quality of speech signals on telephone networks has lead to the development of several algorithms to estimate the subjective quality of a speech signal [9], understood as the quality perceived by a given user. The recommendation P.563 of the International Telecommunications Union (ITU) [6] is an estimation method of the subjective speech quality which includes the effects of the majority of existing impairments in modern telephony networks. Its output is computed from 51 parameters, which are indicators of different possible degradations. The quality measures from this study are mainly based on degradation indicators found in ITU P.563 as well as other work in the literature [10].

According to previous work in the literature [6,8], we define a quality measure as a scalar magnitude which predicts the performance of a given biometric system.

Under such a definition, utterances with poor quality are more likely to be misclassified than those of good quality. A quality measure is defined to be bounded in the range between 0 and 1, where 0 corresponds to the worst possible quality value and 1 to the best one. As this scalar is based on parameters which, in general, do not belong to this range, a mapping function has to be applied, in such a way that for every possible value of a degradation indicator x, the mapping assigns a quality value $Q(x) \in [0,1]$.

The evaluation of quality measures is carried out following the recommendations given by NIST [5], according to which a quality measure is considered useful if as we reject scores with the lowest quality values, the system performance improves.

2.1 Classical quality measures

Quality measures defined in this section have been used before with the purpose of evaluating speech degradation [6,10].

Signal to Noise Ratio (SNR). The SNR degradation indicator has been calculated as follows: making use of a energy-based voice activity detector, each utterance is separated in non overlapping voiced and un-voiced frames of 20 ms. Then, average energy is calculated for both types of frames. Finally, SNR is computed as:

$$\mathrm{SNR} = \log\left(\frac{E_{\mathrm{voiced}}}{E_{\mathrm{unvoiced}}}\right).$$

where E_{voiced} and E_{unvoiced} are the mean energies of the voiced and unvoiced sections. This method for measuring SNR has one main drawback: as it depends on the VAD accuracy, it may have problems to differentiate voiced from un-voiced sections for noisy or very high activity utterances.

We defined the SNR quality mapping function as:

$$Q_{\mathrm{SNR}}(x) = \frac{x}{60}.$$

where x is the SNR value, which is supposed to belong to the range 0-60 dB. Values outside this range will be limited prior to mapping to quality.

Kurtosis LPC (KLPC). Kurtosis is a 4th order statistic which measures the degree of fat tails of a distribution. In this case, kurtosis is applied to the LPC coefficients distribution, as is done in ITU P.563 recommendation [6]. For every 20 ms frame, 21 LPC coefficients are obtained. Then, kurtosis is calculated as:

$$\mathrm{k} = \frac{1}{\mathrm{P}} \sum_{\mathrm{p}=1}^{\mathrm{P}} \left(\frac{\mathrm{a_p} - \frac{1}{\mathrm{P}} \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{a_p}}{\sigma} \right)^4 .$$

where σ represents the standard deviation of LPC coefficients, $\mathrm{a_p}$. Finally, all kurtosis values from all the voiced frames are averaged.

As it will be shown later, the system performance decreases as KLPC increases. According to this, we defined its mapping function as:

$$\mathrm{Q_{KLPC}(x)} = 1 - \left(\frac{\mathrm{x}-3}{8} \right).$$

where x is the KLPC value, which based on our experiments, is supposed to belong to the range 3-11.

ITU P.563 recommendation (P.563). ITU provides an implementation of the algorithm defined on this recommendation. The algorithm generates a Mean Opinion Score (MOS) [11] for each utterance, which is representative of the utterance subjective quality. The MOS belongs to the range 1-5, where 1 corresponds to the worst possible quality value, and 5 to the best one. The input utterance must have a length between 3 and 20 seconds. All utterances duration were between 2 and 5 minutes long, so they had to be divided in smaller fragments and their MOSs were averaged.
The mapping function has been defined according to the MOS scale:

$$Q_{P563}(x) = \frac{(x-1)}{4}. \qquad (5)$$

3 UBML: a novel quality measure for speaker verification

In this work we propose a degradation indicator in the context of speaker verification based on Gaussian Mixture Models (GMM) [13], although the approach can be used in any possible system, no matter the modeling scheme. The proposed measure is motivated by a simple idea. Given that a Universal Background Model (UBM) from a GMM represents the common distribution of speaker features for a given expected operational database, degraded signals are more likely to differ from a UBM than non-degraded signals. Thus, the likelihood between any utterance and the UBM can be used as a measure of speech degradation. Moreover, it is well-

known that speech utterances not matching a given UBM in a GMM system will tend to perform poorly, and therefore a simple measure of the match between a given speech utterance and the UBM like UBML will predict performance for any utterance. Although it may be argued that the likelihood with respect to a UBM may represent many other speaker-dependent information non related to speech degradation, experiments with UBML showed a strong relationship between system performance and this indicator, supporting the assumed hypothesis. In section 5, the validity of this measure is further discussed.

Obtaining UBM likelihood is a mandatory step when using a GMM system, and therefore if such a system is used, the obtention of UBML indicators is costless. However, for other systems UBML can be previously computed and its quality measure used as well. Given a speaker GMM model λ_t and any utterance O for which feature vectors have been extracted, a similarity score is typically computed as:

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

where $p(.\,, \lambda)$ is the probability density function for any model λ. The last term gives the likelihood between any utterance and the UBM:

$$\mathrm{UBML} = \log p(O, \lambda_{UBM})\,.$$

We define the mapping function based on the typical distribution of UBML according to the experiments performed in this work, whose values lay within the range (-13,-5). It is expected that for a given GMM system configuration this value will not significantly change its range among databases. Thus, we map the quality measure as follows:

$$Q_{UBML}(x) = \frac{(x+13)}{8}\,.$$

4 Experiments

4.1 Databases, systems and protocols

In order to evaluate the utility of quality measures, we have used telephone databases and protocols from NIST Speaker Recognition Evaluations 2006 and 2008, which represents a real challenge in terms on session variability [3]. We have

selected both corpuses for experiments in order to show the general behavior of the proposed quality measures among different telephone databases. This fact allows a general strategy of training quality mappings from degradation indicators using a given database (namely NIST SRE 2006) and using such mapping on a different one (namely NIST SRE 2008). For NIST SRE 2008, we have selected the telephone-only subtask of the core condition, namely *short2- short3 tlf-tlf*. For NIST SRE 2006, the whole core condition is used, namely *1conv4w-1conv4w*. For both conditions in the different evaluations, each conversation (coined *short2* for training and *short3* for testing) has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Variability due to different transmission channels, languages and environmental conditions is present, but even more accused in SRE 2008. Although there are speakers of both genders in the corpus, no cross-gender trials are defined.

For score computation, the ATVS GMM system has been used, where speech data known to come from a given speaker is represented using Gaussian Mixture Models adapted from a Universal Background Model. The front-end consists of the extraction of 19 MFCC plus deltas, and processed with rasta filtering and feature warping. Channel factors at feature level have been used for channel compensation [1]. GMM of 1024 mixtures have been used for modeling. Finally, T-Norm has been used for score normalization. The background set for T-Norm cohorts, channel compensation and background modeling is a subset of databases from previous NIST SRE.

4.2 Degradation indicators evaluation

Experiments presented in this section were carried out for 12 different degradation indicators, from previous work in the literature [6,7,10] . They were intended to show the variations of the system performance depending on the magnitude of each indicator, which is useful in order to determine the mapping function from indicator to quality measure. From the whole set of 12, we selected those which showed a clearer relationship with the system performance, namely SNR, ITU P.563 and KLPC.
The experiment was carried out as follows:

8. For every utterance in the databases, each degradation indicator was computed.
9. Scores from the experimental set-up were computed for the described protocols and using the ATVS GMM system.
10. For every score *i*, a mean degradation indicator μ_i is generated computing the arithmetic mean of the indicators for the training utterance and the test segment.
11. Scores are arranged according to their mean degradation indicator μ_i.
12. The first 20% of ordered scores are selected. This is known as set *k.* For each score set *k*, the EER_k is computed, as well as the mean degradation indicator.
13. The last step is repeated 100 times for each set of scores *k*=1,…100. Each time selecting a set of scores with higher degradation indicator. The last set will correspond to the 20% scores with highest degradation indicator.

14. As a result, we obtain 100 EER values and 100 mean degradation values, which correspond to 100 overlapped sets of scores. EER is then represented with respect to its corresponding mean set degradation value.

The following plots show the result of the best performed degradation indicators from the 12 analyzed. We also show the results for the proposed UBML.

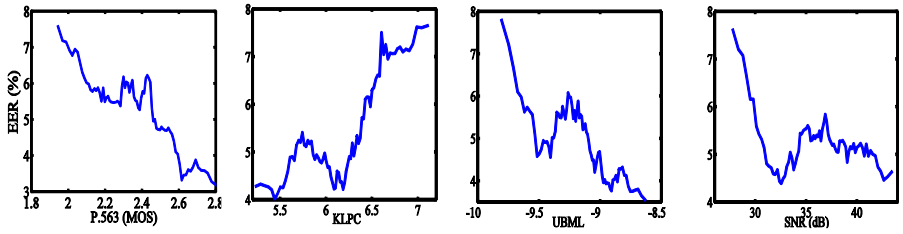


Fig. 1 . EER (%) for every set of scores with a given mean indicator value: P563, KLPC, UBML and SNR, for the NIST SRE 2006 database.

As we can observe, all of them show a clear relationship with the system performance, particularly UBML and P563, for which the EER decreases to 50% for the set of scores with highest qualities.

4.3 Correlation experiments

Given any two quality measures, the linear correlation coefficient among them gives an estimate of how similar is the information they provide about speech degradation in each utterance. This may be interesting in order to combine different quality measures and to optimize the available information to discriminate degraded quality samples. On the following tables we show the correlation coefficients for the five quality measures for both SRE 2006 and 2008 databases.

Table 1. Correlation coefficients for the four quality measures: snr, klpc, p563 and ubml.

	SRE 2006			**SRE 2008**		
	snr	klpc	ubml	snr	klpc	ubml
p563	0.136	0.192	0.223	-0.005	0.145	0.097
snr		0.182	0.386		-0.132	0.536
klpc			-0.034			-0.281

As we can observe, in general all correlation values are moderate. It can be observed a remarkable correlation between UBML and the measures P.563 and SNR. Since P.563 and SNR are well-known degradation indicators, this fact confirms the hypothesis stated in Section 4: UBML is an indicator of signal degradation.

SNR measure presents a low correlation with P.563. This may be due to the low noise level of both databases, since P.563, which selects the strongest of several degradation indicators, is not considering SNR a dominant one. However, SNR has a

clear correlation with UBML, which means that the likelihood between any utterance and the UBM is quite sensitive to the noise contained in the utterance.

4.4 Utility experiments

In this section we try to show the effectiveness of the quality measures as predictors of the system performance. We make use of two kinds of graphic representations: scores-vs-quality scatter plots and error-vs-rejection plots. On the first one, we represent the similarity score against their corresponding quality values (Q), which are obtained combining the qualities of the two involved utterances as:

$$Q = \sqrt{Q_t \cdot Q_{tr}}.$$

where Q_t and Q_{tr} are the quality measures of the test and train utterances.

Since better quality values are supposed to predict better results, target and non-target scores should get more separated as Q gets more close to 1. Regression lines fitted on the plots are intended to show this tendency.

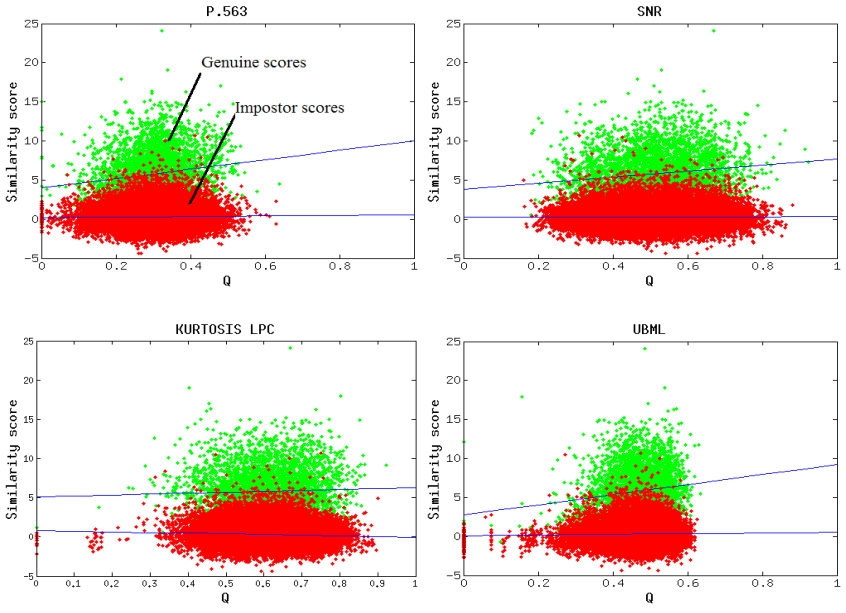


Fig. 2 . Similarity scores against Q, for every quality measure for the SRE 2008 database.

As we can observe, for the quality measures P563, SNR and UBML, scores show a clear tendency to get separated for higher values of Q.

Finally, EER vs reject plots are used as recommended by NIST to show the utility of quality measures [5]. In these plots, the EER is represented against a given

percentage of scores rejected with lowest quality values. The curve is supposed to decrease if the quality measure is useful as the rejection percentage increases. We have represented the results for the rejection fractions: 5, 10, 15, 20 and 25%.

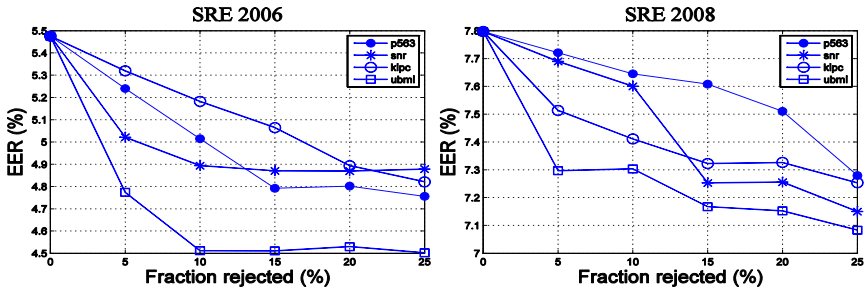


Fig. 3 . EER (%) against rejected scores (%), for both NIST SRE 2006 and 2008 databases.

We can observe that EER decreases for all the quality algorithms as we reject scores. In general, all measures perform better for the 2006 database. It is worth noting that UBML is the best performed measure for both databases, especially for 2006, where the EER decreases a 20% after rejecting the 10% of the scores.

5 Conclusions

In this paper we have analyzed the utility of several quality measures obtained from different indicators of speech degradation typically used in speech processing, namely ITU P.563 estimator of subjective quality, signal to noise ratio (SNR) and LPC Kurtosis (KLPC). We have also proposed a novel quality measure based on the likelihood of a speech segment with respect to a universal model (UBML), which measures degradation in a speech segment by its divergence with respect to such a model. Performance of the quality measures has been presented following the recommendations by NIST, and also using different databases and protocols from NIST Speaker Recognition Evaluations. In all cases, a remarkable utility has been obtained, and a moderate correlation has been observed among different quality measures. Thus, we can argue that the analyzed measures are predictors of speaker verification performance, and therefore they can be used as information in order to compensate for performance drops due to speech degradation.

Future work is mainly related with the use of the obtained quality measures for improving speaker verification performance, and also as complementary information to other data-driven approaches for session variability compensation or fusion in speaker recognition. The potential uses of the promising UBML-based quality measure will be also explored in depth. Finally, a more complete classification of quality measures for speaker verification will be also addressed, including the utility analysis of other different quality measures.

6 Acknowledgements

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. The authors thank Fernando Alonso-Fernandez and Ignacio Lopez-Moreno for fruitful discussions and suggestions.

References

1. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 16(5), pp. 980–988 (2008)
2. Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D., Matějka, P., Schwarz, P., Strasheim, A.: Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Transactions on Audio, Speech and Language Processing, vol. 15(7), pp. 2072–2084 (2007)
3. Przybocki, M.A., Martin, A.F., Le, A.N.: NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006. IEEE Transactions on Audio, Speech and Language Processing, vol. 15(7), pp. 1951–1959 (2007)
4. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J. J.. Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. In Proc. Interspeech 2008, vol. 1, pp. 1493-1496 (2008)
5. Grother, P., Tabassi, E.: Performance of Biometric Quality Measures. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29(4), pp. 531–543 (2007)
6. Malfait, L., Berger, J., Kastner, M.: P.563-The ITU-T Standard for Single-Ended Speech Quality Assessment. IEEE Trans. On audio, speech and language processing, vol. 14, no. 6
7. Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J, Ortega-Garcia, J.: Using Quality Measures for Multilevel Speaker Recognition. Computer Speech and Language, vol. 20(2,3), pp. 192–209 (2006)
8. Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Fronthaler, H., Kollreider, K., Bigun, J.: A comparative study of fingerprint image-quality estimation methods. IEEE Trans. on Information Forensics and Security, vol. 2(4), pp. 734–743 (2007)
9. Grancharov, V., Kleijn, W. B.: Speech Quality Assessment. Springer Handbook of Speech Processing. ISBN 978-3-540-49125-5. Springer Berlin Heidelberg (2008)
10. Richiardi, J., Drygajlo, A.: Evaluation of speech quality measures for the purpose of speaker verification. In Proc. of Odyssey 2008, the ISCA Speaker and Language Recognition Workshop. Stellenbosch, South Africa (2008)
11. Mean opinion score (MOS) terminology, ITU-T Rec. P.800.1, (2003)
12. Reynolds D.A.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, vol. 10, pp. 19--41 (2000)

PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (Software incluido)..... ..... 2.000 €
- Alquiler de impresora láser durante 6 meses ..... 50 €
- Material de oficina ..... 150 €
- Total de ejecución material..... 2.200 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 352 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 132 €

4) Honorarios Proyecto

- 1000 horas a 15 € / hora..... 1500 €

5) Material fungible

- Gastos de impresión ..... 100 €
- Encuadernación ..... 30 €

6) Subtotal del presupuesto

- Subtotal Presupuesto..... 17560 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto..... 2809,6 €

8) Total presupuesto

- Total Presupuesto ..... 19319,6 €

Madrid, Febrero de 2010

El Ingeniero Jefe de Proyecto

Fdo.: Alberto Harriero Castro

Ingeniero Superior de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un estudio de fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz.
En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la

misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.