Universidad Autónoma de Madrid
Escuela politécnica superior

Proyecto fin de carrera

DETECCIÓN AUTOMÁTICA DE VOZ DEGRADADA USANDO MEDIDAS DE CALIDAD

Ingeniería de Telecomunicación

Pedro Cerame Lardies
Junio 2014

DETECCIÓN AUTOMÁTICA DE VOZ DEGRADADA USANDO MEDIDAS DE CALIDAD

AUTOR: Pedro Cerame Lardies
TUTOR: Daniel Ramos Castro

Área de Tratamiento de Voz y Señales
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2014

Resumen

Resumen

En este proyecto se presenta el estudio e implementación de un sistema de detección de voz degradada haciendo uso de distintas medidas de calidad y, posteriormente, se evalúa el impacto de utilizar dichas medidas de calidad como parte del detector de actividad de voz de un sistema de reconocimiento de locutor.

Al comienzo de este proyecto se hace uso de tres medidas de calidad distintas, ya analizadas en otros estudios, para obtener, mediante la combinación de dichas medidas, un único valor que permita, mediante un análisis previo, determinar la elegibilidad de una muestra de voz concreta.

Finalizada la fase de desarrollo del sistema se realiza el experimento de combinar dichos valores con los utilizados por un detector de actividad de un sistema de reconocimiento de locutor desarrollado por el ATVS – *Grupo de Reconocimiento Biométrico*. Tras la realización de este proceso se evalúa el impacto que tienen las medidas de calidad estudiadas sobre el rendimiento total del sistema. Todos los experimentos se han probado sobre una base de datos proporcionada por el NIST – *National Institute of Standards and Technology* (NIST SRE 2012) utilizadas comúnmente en múltiples estudios del estado del arte.

Por último, se presentan las conclusiones y se proponen varias líneas de trabajo futuro.

Palabras Clave

Sistema de reconocimiento de locutor, calidad, indicador de degradación, P.563, SNR, KLPC, rendimiento.

Abstract

In this work we present the study and implementation of a degraded voice detector making use of different quality measures. Also this work evaluates the impact of using these quality measures as part of a voice activity detector used on a speaker recognition system.

At the beginning of this work we use three different quality measures, already analyzed in other studies, and obtaining, by the combination of these measures, one value that permits determinate the eligibility of a voice sample.

When the developing phase is done, the quality measures are combined with the labels of a voice activity detector, developed by the ATVS group. After that, we evaluate the impact of these quality measures on the speaker recognition system. The database used for the experiments is the NIST SRE 2012.

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

Key words

Speaker recognition system, quality, degradation indicator, P.563, SNR, KLPC, performance.

Agradecimientos

En primer lugar, quería agradecer a mi tutor, Daniel Ramos, por darme la oportunidad de realizar este proyecto, por guiarme durante este (largo) proceso y animarme a seguir y por el esfuerzo que me ha dedicado. Además, me gustaría agradecer también a todos los profesores que he tenido, no sólo en la universidad sino también en el colegio, pues sin ellos no habría podido adquirir los conocimientos necesarios para completar esta etapa de mi vida.

A todos mis amigos, porque han estado durante los momentos buenos y los menos buenos y, aunque no puedo nombrar a todos, me gustaría agradecérselo en especial a Alex, porque ha estado conmigo casi desde que puedo acordarme. Gracias por hacerte notar incluso a pesar de los kilómetros que últimamente nos separan.

También a mis compañeros y amigos de la universidad, por que hemos pasado muy buenos momentos tanto dentro como fuera de la facultad. Han sido unos años de los que me acordaré siempre.

En especial quiero agradecer el apoyo de mi familia. Más concretamente a mis padres, porque me han apoyado en todo momento. Gracias por ayudarme a valorar mis decisiones y enseñarme a ser coherente con ellas. Gracias por no permitirme desfallecer y por animarme a levantarme.

Por último quería dar las gracias a Marta, por escucharme en los momentos que más lo necesitaba, por ayudarme a no rendirme. Has sido la fuerza que me animó a seguir.

Índice general

Índice de figuras

Índice de cuadros

1 Introducción

1.1. Motivación del proyecto

En los últimos años, el uso de sistemas de reconocimiento biométrico ha ganando mucho peso en casi todos los ámbitos, debido a que proporciona un sistema de identificación fiable y con una menor probabilidad de falsificación con respecto a otros métodos más extendidos.

Un sistema biométrico es, en esencia, un sistema de reconocimiento de patrones cuya función consiste en obtener la información biométrica de un individuo, extraer cierta característica de esta información y compararla con un patrón específico definido en una base de datos. Dependiendo del contexto sobre el que se aplique, un sistema biométrico puede operar tanto en modo de verificación como en modo de identificación. [1] En el caso de la **verificación**, el sistema debe comprobar que el usuario que se está identificando es quien dice ser mientras que el modo de **identificación**, consiste en determinar la identidad del usuario en cuestión.

Dentro del ámbito forense, el uso de sistemas automáticos de reconocimiento de locutor se ha visto incrementado en los últimos años. La mejora de la precisión que ha experimentado esta tecnología, unida a un estudio más exhaustivo sobre el rol del reconocimiento automático de locutor en la ciencia forense, son las razones que han permitido este incremento. [2]

Las grabaciones de voz que intervienen durante una investigación criminal son de dos tipos: grabaciones incriminatorias (**dubitadas**) procedentes de micrófonos ocultos, pinchazos en la línea telefónica, etc... y grabaciones hechas al sospechoso en entornos controlados (**indubitadas**). La evidencia forense viene dada por el grado de similitud entre estas dos muestras de voz.

En estos casos se trabaja con muestras grabadas en situaciones muy adversas o en un entorno con ruido ambiente muy variable y que, en muchos casos, pueden degradar la señal de voz e influir en la fidelidad de ésta. Dentro de una misma muestra de voz, la señal puede experimentar muchos cambios de calidad que es posible afecte al proceso de reconocimiento de locutor. Por lo tanto es necesario encontrar esas zonas en las que la calidad es menor y eliminarlas, de forma que disminuya el número de muestras a procesar.

El objetivo de este PFC es el de realizar un sistema que utilice tres métodos de estimación de calidad de la voz ya estudiados [3] para desechar, en función de su calidad, las muestras de la señal de voz que no van a formar parte del proceso de reconocimiento de locutor. Este procedimiento pretende obtener una muestra de voz menos degradada y más fidedigna que permita aumentar el rendimiento del sistema de reconocimiento de locutor.

1.2. Objetivos y enfoque

El objetivo principal de este proyecto es mejorar un detector de actividad de voz automático utilizando varios métodos de estimación de calidad de la voz para, de esta manera, intentar aumentar el rendimiento del proceso de reconocimiento de locutor. Para ello se realizará una revisión del estado del arte en relación a la calidad de voz y su uso en el ámbito de reconocimiento de locutores en situaciones adversas. A continuación se seleccionarán tres métodos de cálculo de la calidad de una señal de voz de los muchos propuestos en la literatura, a saber: *SNR (Signal to Noise Ratio), KLPC (Kurtosis LPC) y P563*. A diferencia de los demás trabajos previos, en este proyecto se generarán algoritmos que permitan calcular dichas medidas de calidad no globalmente para fragmentos completos de la señal de voz, sino de diferente forma en diversas partes de cada grabación.

A la hora de realizar las pruebas se utilizará la base de datos que proporciona el NIST, más concretamente los datos correspondientes al NIST SRE 2012 – *NIST Speaker Recognition Evaluation 2012*, un estudio bianual que trata de implementar mejoras importantes en la tecnología de reconocimiento de locutor y que cuenta con muestras de habla tanto de origen telefónico como microfónico (no han sido transmitidas por una red de telefonía).

El sistema que se va a implementar deberá ser capaz de etiquetar, con sus correspondientes valores de calidad, cada una de las tramas resultantes de un determinado análisis a corto plazo de una muestra de voz. Una vez se conocen los valores de calidad de cada una de las tramas, las muestras que no superen cierto umbral serán desechadas y no se verán involucradas en el proceso de reconocimiento de locutor.

1.3. Metodología y plan de trabajo

La realización de este proyecto se ha dividido en las siguientes etapas:

- **Formación.** Donde se han adquirido los conocimientos básicos sobre reconocimiento de patrones utilizando el libro [4]. El estudio sobre las diferentes métodos de estimación de la calidad de voz han permitido realizar un resumen sobre el estado del arte presentado en el Capítulo 3 de este proyecto.
- **Desarrollo.** Durante esta etapa se han implementado los scripts necesarios para calcular las medidas de calidad trama a trama. Se han propuesto nuevas soluciones para problemas no existentes, debido a que la mayoría de los algoritmos implementados en el laboratorio calculaban la calidad de forma global en una locución, no trama a trama.
- **Experimentación.** Se han realizado una serie de experimentos para evaluar la eficacia de incluir medidas de calidad en la detección de voz para un sistema de reconocimiento de locutor. En esta etapa se han evaluado los resultados obtenidos y se han extraído diversas conclusiones.

- **Escritura de la memoria.** Redacción de este documento durante la realización de las demás tareas.

1.4. Organización de la memoria

- **Capítulo 1. Introducción.** Se exponen los motivos que han impulsado el desarrollo de este proyecto así como los objetivos que se pretenden alcanzar.
- **Capítulo 2. Introducción a la biometría.** Conceptos básicos de biometría (tipos de rasgos, características, etc.) y sistemas biométricos (arquitectura, modos de funcionamiento, evaluación del rendimiento y aplicaciones).
- **Capítulo 3. Calidad.** En este capítulo se describen algunas de las características de la calidad dentro del ámbito de reconocimiento de locutor, tomando como referencia algunos trabajos sobre el estado del arte.
- **Capítulo 4. Cálculo y estudio de las medidas de calidad.** Se realiza el cálculo de las medidas de calidad sobre las locuciones y se evalúa el impacto que tendrá sobre el sistema de reconocimiento utilizado en los experimentos del Capítulo 5.
- **Capítulo 5. Experimentos realizados y resultados.** Se describe el marco experimental y los procedimientos que se van a utilizar para la consecución de los experimentos. Además, se presentan y evalúan los resultados obtenidos con el fin de extraer conclusiones útiles que permitan establecer futuras líneas de investigación.
- **Capítulo 6. Conclusiones y trabajo futuro.** Se presentan las conclusiones extraídas tras el estudio de los resultados y se definen las posibles líneas de trabajo que se han podido extraer durante la realización de este proyecto.

2

Introducción a la biometría

La biometría tiene como finalidad el reconocimiento de individuos de forma automática, haciendo uso de ciertas características o rasgos biométricos que pueden ser tanto físicos como conductuales. [5]

2.1. Características de los rasgos biométricos

Para que un rasgo biométrico sea considerado como tal, debe cumplir ciertos requisitos: [1]

- **Universalidad:** todas las personas deben poseer este rasgo.
- **Unicidad:** debe ser suficientemente único en cada individuo como para permitir diferenciarlos.
- **Estabilidad:** esta característica debe permanecer invariante a lo largo del tiempo
- **Mensurabilidad:** debe poder ser medido cuantitativamente.

Además, en la práctica, es necesario tener en cuenta las siguientes características para que se pueda considerar un sistema de reconocimiento biométrico. Éstas son:

- **Rendimiento:** hace referencia a la velocidad y la precisión con la que trabaja un sistema, así como los factores operacionales y del entorno que le afectan.
- **Aceptabilidad:** los usuarios deben estar dispuestos a utilizar este sistema de reconocimiento. Debe estar socialmente aceptado.
- **Evitabilidad:** es necesario contar con un grado de seguridad alto para que el sistema sea difícil de eludir usando métodos fraudulentos.

2.2. Rasgos biométicos

Existen un gran número de rasgos biométricos que pueden ser usados en aplicaciones de reconocimiento de individuos. Cada rasgo tiene sus fortalezas y sus debilidades, y la elección de uno u otro rasgo dependerás del uso que se le quiera dar. A continuación se presenta un breve resumen de cada uno de estos rasgos:

- **ADN:** el ácido desoxirribonucleico se encuentra presente en cada célula y es prácticamente único para cada individuo, salvo en el caso de los gemelos monocigóticos. Este rasgo es muy utilizado en el ámbito forense pero se ve limitado en el uso de otras aplicaciones de reconocimiento biométrico, pues se trata de un método muy invasivo y que requiere un tiempo de procesado muy alto.
- **Cara:** se trata de un método no invasivo que es posiblemente el más común dentro del campo de reconocimiento biométrico. Los métodos de reconocimiento facial más populares son los basados en el contorno de los rasgos de la cara (ojos, nariz, labios, etc...) y los que hacen uso de un análisis completo de la imagen facial. Estos sistemas presentan algunos problemas a la hora de reconocer imágenes faciales capturadas en condiciones de iluminación muy variables o tomadas desde ángulos muy diferentes.
- **Firma:** el modo en que una persona firma es característico de cada individuo. Este método requiere contacto con el instrumento de escritura y un esfuerzo por parte del usuario. por lo que ha sido bien aceptado en el ámbito gubernamental y a la hora de realizar transacciones comerciales. La firma se trata de un rasgo biométrico de carácter conductual que cambia con el tiempo y se ve influido por las condiciones físicas y emocionales del firmante. Además, una persona especializada en la falsificación podría reproducir una firma que engañe al sistema.
- **Geometría de la mano:** un sistema de reconocimiento basado en este rasgo utiliza varios aspectos extraídos de la mano como pueden ser su contorno, el tamaño de la palma y la longitud o el ancho de los dedos. La técnica es muy simple, relativamente fácil de usar y no demasiado cara.
- **Huella dactilar:** se trata de un rasgo que se lleva usando en el reconocimiento de individuos desde finales del siglo XIX. Este rasgo consta de una gran unicidad y tiene la ventaja de permanecer invariable en el tiempo, dotándole de un gran poder discriminativo. Su mayor utilidad ha sido siempre aplicada al ámbito forense pero actualmente se puede encontrar en un gran número de aplicaciones comerciales.
- **Iris:** es la región anular del ojo que se encuentra entre la pupila y la esclera. La textura compleja del iris proporciona una gran información que facilita el reconocimiento entre individuos. Cada iris es único y, al igual que las huellas dactilares, hasta el iris de los gemelos es distinto para cada uno. Los primeros sistemas de reconocimiento de iris requerían mucha participación por parte del usuario y eran bastante caros. Sin embargo, los últimos sistemas son más cómodos para el usuario y no requieren de una inversión tan costosa.
- **Modo de andar:** a pesar de no parecer un rasgo muy identificativo, el modo de andar de una persona es lo suficientemente discriminatorio como para ser usado en aplicaciones de verificación de baja seguridad.
- **Voz:** la voz se compone de una combinación de rasgos físicos y conductuales. Las características de la voz de un individuo vienen dadas por la fisionomía del sistema que genera el sonido (tracto vocal, labios, boca y cavidad nasal). Estos rasgos cambian debido a la

edad, el estado de ánimo, las condiciones médicas, etc. Un sistema de reconocimiento dependiente de texto se basa en la pronunciación de una frase fijada de antemano, mientras que un sistema independiente de texto reconoce al interlocutor independientemente de lo que diga.

2.3. Sistemas biométricos

2.3.1. Funcionamiento

A grandes rasgos, un sistema biométrico se puede definir como un sistema de reconocimiento de patrones que toma como entrada cierta muestra de un rasgo biométrico de un individuo, y le asigna una identidad determinada. La Figura 2.1 esquematiza las etapas básicas de un sistema biométrico:

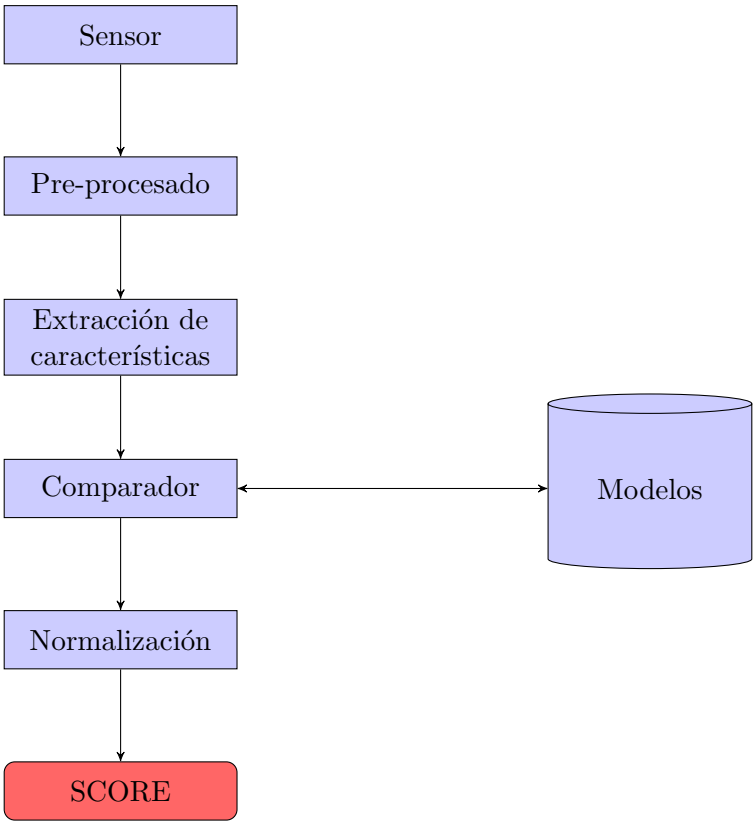

Figura 2.1: Esquema del funcionamiento de un sistema biométrico.

En un primer lugar, se adquiere la información biométrica del individuo utilizando un sensor. Se trata de una etapa muy importante pues permitirá, en las siguientes etapas, extraer la mayor cantidad de información posible sobre la característica biométrica que se esta midiendo. Una vez registrada la muestra, es necesario realizar un procesado previo para eliminar posibles ruidos o distorsiones facilitando así la extracción de la información en la siguiente etapa. Esta misma información permitirá seleccionar una serie de parámetros o características discriminantes. Dichos parámetros se comparan con uno o varios modelos, produciendo una puntuación de similitud (score) entre cada uno de los modelos y la muestra, proporcionando así una medida cuantitativa del parecido entre ambas muestras.

2.3.2. Modos de funcionamiento de un sistema biométrico

Un sistema biométrico puede operar en dos modos diferentes:

- **Modo de verificación o detección:** el individuo afirma poseer cierta identidad. La finalidad de este sistema es tratar de demostrar si esa identidad es realmente suya o no. Para ello, el usuario debe proporcionar su rasgo biométrico y su identificación. Se realiza una comparación con el modelo correspondiente a la identificación proporcionada y se obtiene una puntuación. Con dicha puntuación, y teniendo en cuenta cierto umbral, el sistema decidirá si acepta o rechaza al usuario.

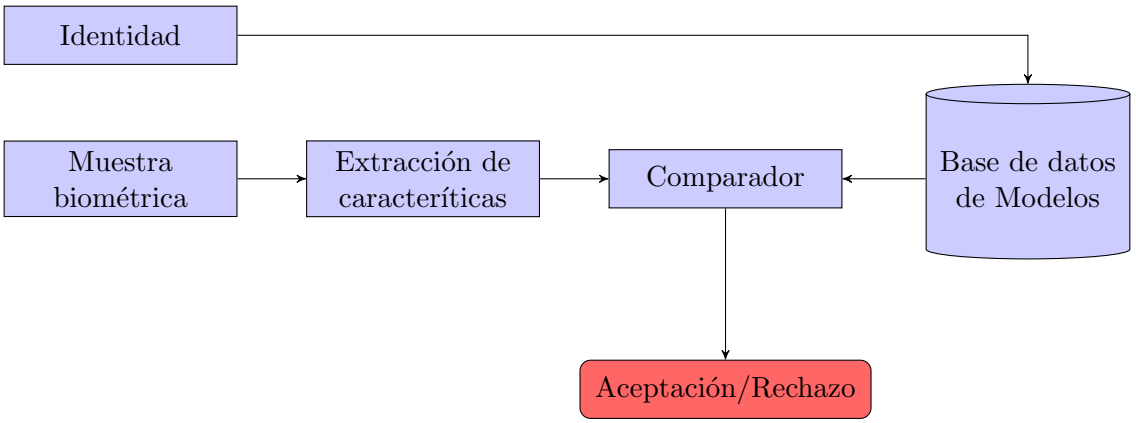

Figura 2.2: Esquema del funcionamiento de un sistema en modo de verificación.

- **Modo de identificación:** en este modo, el usuario proporciona un rasgo biométrico y el sistema determina si el individuo se encuentra en la base de datos (identificación) o no (rechazo). Este modo requiere un coste computacional mucho más elevado pues será necesario comparar el rasgo proporcionado con todos los modelos de la base de datos. Tras esta comparación se obtienen una serie de puntuaciones para cada modelo de la base de datos y, salvo que ninguna de las puntuaciones alcance el umbral permitido, en cuyo caso el usuario será rechazado, se elegirá la muestra cuya puntuación sea mayor. Además, dentro de este modo se pueden distinguir dos tipos:

  - *Conjunto cerrado:* el usuario siempre pertenecerá a algún modelo de los almacenados en el sistema.
  - *Conjunto abierto:* es posible que el usuario no pertenezca a ninguno de los modelos existentes en la base de datos.

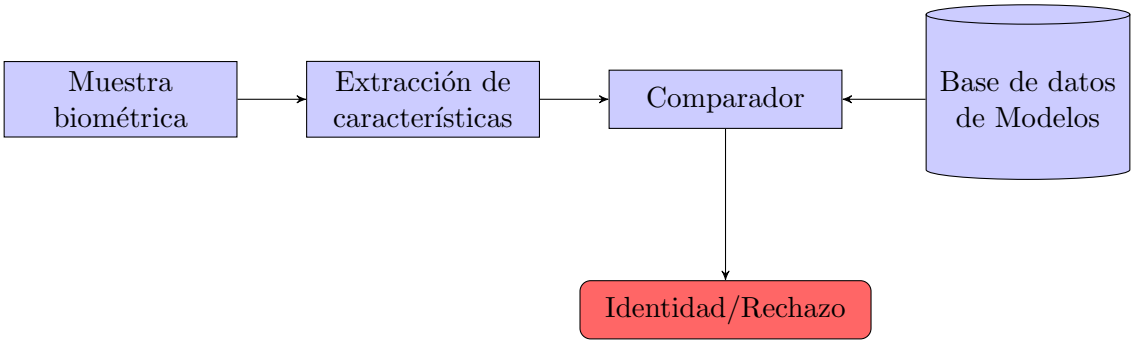

Figura 2.3: Esquema del funcionamiento de un sistema en modo de identificación.

2.3.3. Problemas y limitaciones de los sistemas biométricos

El rendimiento de los sistemas biométricos se puede ver limitado por una serie de factores que se citan a continuación:

- **Actitud:** La actitud del individuo que esta usando el sistema puede dar lugar a una mala medición de la muestra en cuestión, pudiendo no reflejarse correctamente las características de dicho individuo.
- **Problemas en el proceso de extracción:** El entorno en el que se esta realizando la extracción o la precisión del dispositivo con el que se realiza la captura de la muestra son algunos factores que pueden generar problemas durante esta parte del proceso.
- **Envejecimiento o alteraciones:** Uno de los problemas mas comunes viene dado por el envejecimiento de los rasgos biométricos del individuo. En este caso, la voz es uno de los rasgos mas afectados por este problema. Las alteraciones del rasgo (cicatrices, tatuajes, afonías...) pueden también dificultar la tarea de reconocimiento.
- **Variabilidad entre sesiones:** Originada por diversos factores como la distorsión, la calidad, etc.

2.3.4. Aplicaciones de los sistemas biométricos

Establecer la identidad de un individuo se ha convertido en una necesidad critica en la sociedad. Esta necesidad de autenticación ha incrementado el uso de sistemas de reconocimiento biométrico en ámbitos muy distintos. Sus aplicaciones se pueden dividir en tres grandes bloques [6]:

1. **Comerciales.** Protección de datos, acceso a internet, e-commerce, uso de cajeros automáticos o tarjetas de crédito, teléfonos móviles, etc.
2. **Gubernamentales.** Documento de identidad, licencia de conducir, Seguridad Social, pasaporte, etc.
3. **Forenses.** Identificación de cuerpos, investigación criminal, evaluación de evidencias, pruebas de paternidad, etc.

2.3.5. Aceptación en la sociedad y privacidad

La voz, dentro de los rasgos biométricos utilizados en la actualidad, es, incluso a pesar de sus limitaciones, uno de los más aceptados en la sociedad. Existen dos principales razones que sustentan esta idea [7]:

1. **Aceptado por la sociedad** de forma común. No supone demasiado esfuerzo ni una amenaza para la privacidad el hecho de usar la voz para identificarse, pues se trata de un acto muy cotidiano.
2. Gracias a la **red telefónica** es posible utilizar esta clase de sistemas desde casi cualquier punto del planeta de forma remota.

2.4. Sistemas de reconocimiento de locutor

2.4.1. Información de la identidad en la señal de voz

A la hora de realizar un sistema de reconocimiento de locutor, es necesario tener en cuenta que el proceso de producción de la voz es muy complejo y viene determinado por las características físicas del individuo (el denominado *tracto vocal*) y una serie de factores tales como la educación, acento, contexto social, estado anímico y de salud, etc. Un sistema de reconocimiento se fundamenta en la idea de que, la percepción humana se basa en esos factores anteriormente citados para reconocer a una persona por su voz. Para ello, un sistema automático extrae la información de la señal de voz a distintos niveles, como por ejemplo [8]:

- **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. Se usan ventanas de muy corta duración (milisegundos) considerándose que, en ese tiempo, la configuración del tracto vocal permanece invariable. A partir de estas ventanas se extraen una serie de parámetros.
- **Nivel fonético.** Una de las formas de discriminar una voz es basándose en las características fonéticas y la pronunciación del individuo, pues cada persona realiza un uso de los fonemas y las sílabas diferente.
- **Nivel prosódico.** La prosodia estudia aquellos elementos de la expresión oral tales como la entonación, los tonos, la energía de la señal, etc. El papel que estos elementos juegan dentro de la producción de las palabras se asocian a una serie de variaciones de la frecuencia fundamental, la duración y la intensidad que constituyen los parámetros prosódicos.

2.4.2. Funcionamiento de un sistema de reconocimiento de locutor

Cualquier sistema de reconocimiento de locutor, independientemente de su aplicación, sigue una misma estructura fundamental. Esta estructura se compone de dos fases:

1. **Entrenamiento:** es necesario generar un modelo representativo de la identidad de cada locutor. Para realizar este entrenamiento será necesario contar con una o varias muestras de ese locutor en cuestión.

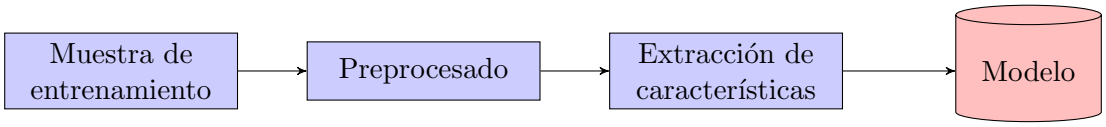

Figura 2.4: Esquema del funcionamiento de un sistema de reconocimiento de locutor durante la fase de entrenamiento.

2. **Cálculo de la similitud:** para realizar este cálculo será necesario comparar una locución con el modelo estadístico correspondiente a una identidad concreta. Tras esta comparación se obtiene una puntuación (*score*) que indica el grado de similitud entre la locución y el modelo.

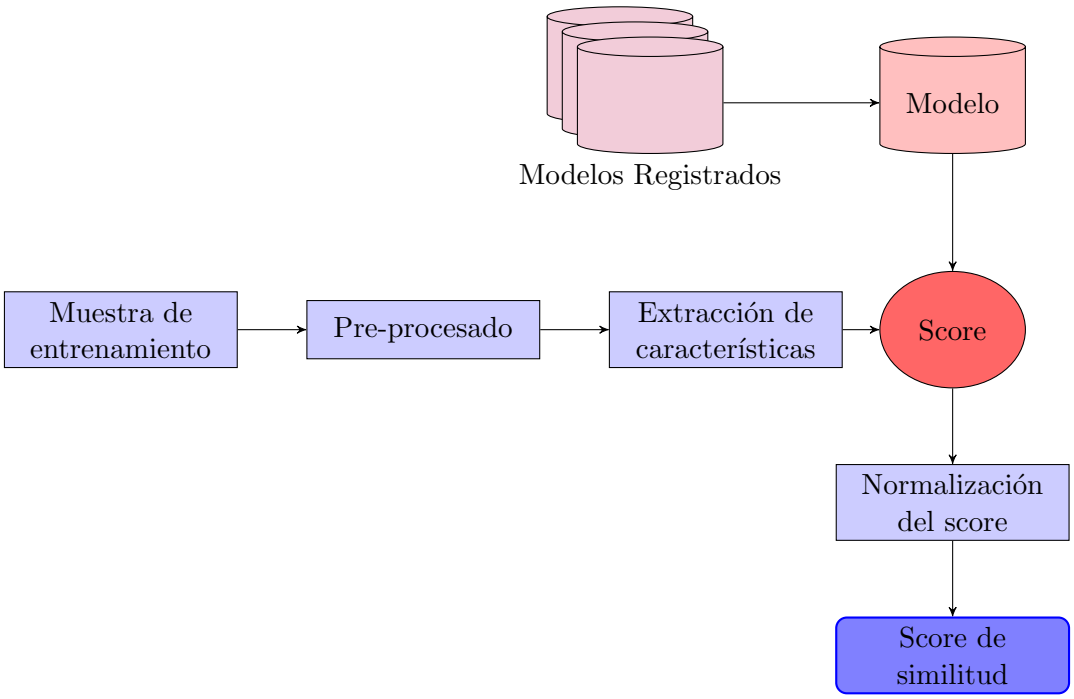

Figura 2.5: Esquema del funcionamiento de un sistema de reconocimiento de locutor durante la fase de entrenamiento.

Dependiendo del modo de funcionamiento del sistema, esta puntuación se tratará de una manera u otra:

- Si hablamos de un sistema que funciona en modo de *verificación*, se establecerá un umbral para determinar si la locución y el modelo pertenecen al mismo locutor.
- En el caso de un sistema en modo de *identificación*, se realizarán N comparaciones (N = número de identidades que pueden asignarse a un locutor). Una vez obtenidas las N puntuaciones, se tomará una decisión.
- Recientemente, se entrenan dos modelos (*ivectors*) y se comparan entre ellos. Desaparece la distinción entre modelo y test, y se comparan únicamente dos modelos. Este tipo de sistema se usará en este proyecto.

2.4.3. Rendimiento

Para poder poner en funcionamiento un sistema de reconocimiento de locutor de forma efectiva es necesario utilizar algún mecanismo de evaluación que proporcione una medida del rendimiento del sistema. En sistemas de verificación, se utilizan más comúnmente métodos que midan la capacidad de discriminación de dicho sistema.
Como se ha explicado en puntos anteriores, a la salida de un sistema de reconocimiento de locutor se obtiene una puntuación, una medida sobre el grado de similitud entre la muestra y el modelo. Con esta puntuación, el sistema debe establecer un umbral para decidir si el usuario es quien dice ser (*genuino*) o no (*impostor*). En este punto se pueden cometer dos tipos de errores:

- **Falso Rechazo.** Se denomina así al error en el que se incurre cuando el sistema detecta a un usuario impostor pero realmente se trata de un usuario genuino.
- **Falsa Aceptación.** En este caso, el sistema detecta un usuario genuino pero en realidad se trata de un usuario impostor.

En términos de verificación, es muy común utilizar las denominadas curvas DET - *Detection Error Tradeoff* para representar estos dos tipos de errores uno frente a otro en un eje normalizado. De esta forma se representa una única curva definida para todos los posibles puntos de trabajo del sistema a evaluar. En este punto se puede calcular el EER (*Equal Error Rate*) ya que se trata del punto en el que la curva DET corta con la bisectriz de la gráfica. Utilizando este tipo de curvas es muy fácil visualizar el rendimiento de un sistema, pues cuanto más se acerque la curva DET al origen, mayor poder de discriminación tendrá el sistema [9].

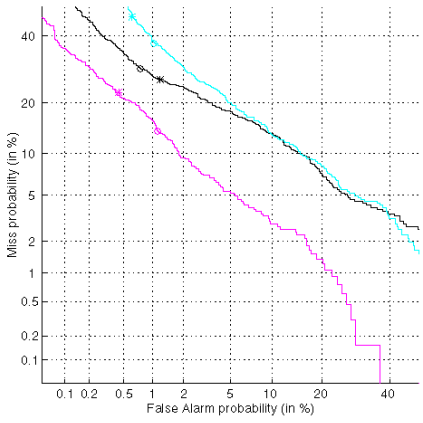

Figura 2.6: Ejemplo de curva DET. Figura extraída de [9].

3
Calidad

En el siguiente capítulo se van a explicar algunos de las características más representativas de la calidad dentro del ámbito del reconocimiento biométrico, así como los tipos de medidas de calidad de voz que se usarán en este proyecto. Para ello se tomarán como referencia distintos trabajos del estado del arte.

3.1. Introducción

3.1.1. Definición de calidad

La primera pregunta que se puede plantear cuando se habla de calidad es, básicamente, cuál es el significado de calidad dentro del ámbito de la biometría. En este caso, se puede decir que una muestra presenta buena calidad si es apropiada para su uso en un sistema de reconocimiento. Este punto de vista puede distar de la concepción de calidad del ser humano. Por ejemplo, si una persona ve una huella dactilar que parece nítida, con poco ruido y un buen contraste podría decir de manera razonable que se trata de una muestra con buena calidad. Sin embargo, si la imagen contiene un bajo número de minucias, un sistema de reconocimiento basado en minucias no trabajará de forma óptima [10].

La calidad de una muestra biométrica se puede considerar desde tres puntos de vista diferentes [11]:

1. **Carácter.** Se refiere a la calidad que se atribuye a las características físicas inherentes a cada sujeto en cuestión.
2. **Fidelidad.** Grado de similitud entre una muestra biométrica y su fuente. Esta fidelidad puede determinarse por dos factores:
   - Fidelidad de adquisición.
   - Fidelidad de procesado.
   - Fidelidad de extracción.

   La figura 3.1 ilustra esta clasificación.

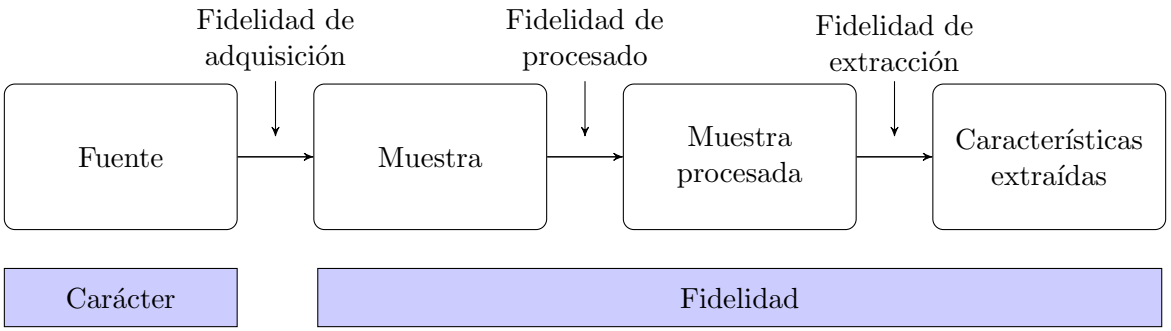

Figura 3.1: Esquema de los factores que influyen en la calidad de una muestra biométrica.

3. **Utilidad.** Se refiere al impacto que tiene una muestra biométrica individual sobre el rendimiento total de un sistema biométrico.

3.1.2. Factores que influyen en la calidad biométrica

Existen un número de factores que afectan a la calidad de una muestra biométrica. Los diferentes factores que tienen impacto sobre la calidad de la muestras pueden clasificar dependiendo de su relación con las diferentes partes del sistema. Atendiendo a esta clasificación, podemos encontrar cuatro diferentes factores [12]:

1. **Factores relativos al usuario.** Son los factores relacionados con la **fisiología** y el **comportamiento** del usuario.

   - *Factores fisiológicos.* Son los factores más difíciles de controlar, puesto que dependen enteramente del sujeto. Algunos de estos factores no afectan directamente a la degradación de la muestra pero sí a su variabilidad. Esta variabilidad, si no se tiene en cuenta durante el reconocimiento, puede afectar al rendimiento del sistema. Otros factores, como las heridas o las enfermedades, pueden alterar los rasgos biométricos incrementando su variabilidad.
   - *Factores de comportamiento.* Estos factores son más fáciles de paliar que los relacionados con la fisionomía del individuo. Dependen del estado de ánimo del usuario, pues este puede no tener ganas de proporcionar su característica biométrica, estar nervioso, distraído o cansado. En muchas ocasiones es posible corregir estos factores, pero no siempre es lo más apropiado, pues se trata de cambiar los hábitos de los usuarios.

2. **Factores relativos a la interacción entre el sensor y el usuario.** Son más fáciles de controlar que los factores relativos al usuario aunque este sigue formando parte en ellos. Dentro de esta categoría existen dos tipos de factores: **ambientales** y **operacionales**:

   - *Factores ambientales.* Estos factores pueden amortiguarse si se controla el entorno en el que se toma la muestra. Si se trata de un entorno controlado, es bastante sencillo disminuir este tipo de degradación, pero si se trata de un sensor que se encuentra en un entorno menos controlado, como un ambiente de exteriores, el problema no es tan sencillo de resolver. En este último caso, hará falta tener en cuenta, no sólo los factores ambientales relativos a la adquisición de la muestra sino también los que afectan al sensor en sí (humedad, ruido, iluminación, etc.).

- *Factores operacionales.* Igual que en el caso de los factores ambientales, los operacionales pueden ser controlados si tenemos influencia sobre el acto de adquisición de la muestra en sí. Un factor importante que tiene que ver con la operatividad del sistema es el relativo al tiempo que pasa entre adquisiciones (*envejecimiento*). Este factor se basa en la premisa de que la información extraída de un individuo en dos momentos diferentes puede variar. Algunos sistemas son más sensibles a estos cambios que otros tales como la firma o la voz.

3. **Factores relativos al sensor.** Existen varias características del sensor que pueden afectar a la calidad de la muestra biométrica adquirida: *la facilidad de uso, el tamaño del área de adquisición de la muestra, su fiabilidad y resistencia física, su rango dinámico* o *el tiempo que se necesita para adquirir la muestra.* Para paliar estos problemas es necesario que el sensor trabaje con ciertos estándares pues facilitará el reemplazamiento del sensor sin que afecte a la fiabilidad del proceso de adquisición.

4. **Factores relativos al sistema de procesado.** Son los factores más fáciles de controlar, porque están relacionados con el procesamiento de la muestra una vez ha sido adquirido por el sensor. Los factores que afectan en este caso son: *el formato de la información* y *los algoritmos utilizados.*

3.1.3. Implicaciones del uso de datos con mala calidad

Los problemas de calidad con muestras biométricas reducen el rendimiento de la comparación y, en algunos casos extremos, puede hacer que esta comparación sea imposible de realizar. Dependiendo del sistema, los problemas de calidad también pueden ralentizar el proceso de reconocimiento. [13].

3.1.4. Modos de mejorar la calidad de una muestra biométrica

Como se ha podido comprobar, los problemas asociados al uso de muestras con mala calidad pueden ser muy diversos y tener graves consecuencias sobre el rendimiento y la precisión de un sistema de reconocimiento. Es por ello que hay que intentar mejorar la calidad de las muestras obtenidas o, si esto no es posible, disminuir el impacto de dichas muestras sobre el sistema. Existen pues, varias soluciones para mejorar de una manera o de otra la calidad de una muestra biométrica:

- **Recapturación de las muestras [13].** Se trata de evitar que las muestras de mala calidad sean almacenadas en el sistema (para, por ejemplo, la generación de un modelo). Si una muestra es de mala calidad será rechazada y se repetirá la captura del rasgo biométrico cuantas veces haga falta hasta que se asegure un umbral mínimo de calidad.

- **Generación de modelos.** Los modelos generados a partir de las muestras biométricas suelen estar muy influenciados por la calidad de dichas muestras. Esta variabilidad de la calidad se puede producir entre muestras distintas o incluso entre varias zonas de una misma muestra. Toda esa información de peor calidad debería tener menor peso dentro del modelo global para así evitar futuros errores.

- **Comparación del score.** Se trata del mismo método que en la generación de los modelos pero atendiendo a la calidad del score global. En este caso, según la calidad de cada sección de la muestra se le otorgará mayor o menor peso, permitiendo así que, siempre que la muestra sea útil, mejoren los resultados del score.

- **Fusión a partir de la calidad [14].** Consiste en aprovechar que no todos los sistemas se ven afectados de la misma manera ni en la misma medida por la calidad de las muestras recogidas y fusionarlos, dando mayor peso a los sistemas que aparentemente son más robustos frente a un descenso de la calidad.

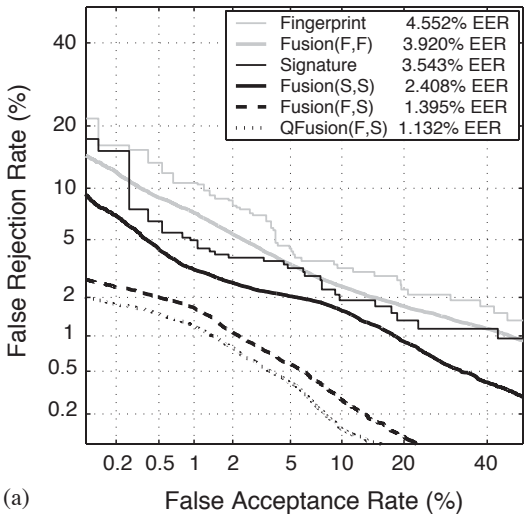

Figura 3.2: Ejemplo de fusión de dos sistemas distintos. Mediante la curav DET se evalúa la mejora de rendimiento al fusionar dos sistemas, en este caso la huella dactilar y la firma. Figura extraída de [14]

- **Normalización de scores a partir de la calidad [12].** Se trata de aplicar diferentes parámetros de normalización de scores en función de la calidad de la muestra.

3.2. Calidad de la voz

El rápido despliegue de las aplicaciones de procesado del habla han incrementado la necesidad de evaluación de la calidad en la voz. El éxito de cualquier tecnología nueva depende en gran parte de la opinión del usuario sobre la calidad de la voz percibida [15].

3.2.1. Medidas de calidad

En este apartado se clasifican los principales tipos de medidas de calidad de voz independientemente de si utilizan en sistemas de reconocimiento o no. Atendiendo al tipo de procedimiento utilizado para su obtención, se han clasificado las medidas de calidad en las siguientes categorías:

- **Modelo de estimación subjetivo**

  Uno de los primeros métodos que se utilizó para medir la calidad de una señal de voz se basó en la observación de la calificación que habían establecido varias personas a ciertas locuciones (*calidad subjetiva*). Es el método que más se ajusta al modelo de la percepción humana.

  La eficacia de estos métodos dio pie a establecer una serie de recomendaciones sobre los experimentos a realizar para medir este tipo de calidad [16]. Existen también medidas de calidad objetivas que tratan de estimar la calidad subjetiva, como la **P.563** (medida recomendada por el ITU - *International Telecommunication Union* [17]).

- **Análisis del ruido**

  Este tipo de análisis pretende realizar una estimación del nivel de ruido de cierta locución. La mayoría de medidas de calidad están ligadas de alguna manera con el análisis de ruido, pero la más común es la que trata de medir la relación que existe entre la señal de voz y el ruido, la **SNR** - *Signal-to-Noise Ratio*. Esta medida es una de las que se ha usado en este proyecto.

- **Análisis de las estadísticas de la voz.**

  Las medidas de esta clase tratan de medir el nivel de degradación de una muestra mediante los estadísticos de la señal. Las dos más comunes se denominan *kurtosis* y *skewness*, y en el ITU-P563 [18] se utilizan aplicadas a los parámetros MFCC - *Mel Frequency Cepstral Coefficients* y LPC - *Linear Predictive Coding* de la locución.

3.3. Medidas de calidad utilizadas

En este proyecto se han utilizado algunas de las medidas de calidad propuestas en [3] que son: *P.563*, *SNR* y *KLPC*.

3.3.1. ITU-P.563

Se trata de una medida que pretende calcular o predecir, de forma objetiva, la calidad subjetiva de una locución.

La ITU provee un algoritmo específico para la implementación de esta medida. Este algoritmo posee una gran complejidad y utiliza diferentes parámetros para estimar el tipo de degradación que predomina en la señal y, además, genera una puntuación MOS - *Mean Opinion Score* [19] que determinará la calidad de dicha locución. La puntuación MOS tendrá un valor que varía de 1 a 5, donde 1 corresponde al peor valor de calidad posible y 5 al mejor.

Las locuciones a evaluar deben tener una longitud temporal de entre 3 y 20 segundos. A diferencia del estudio [3], en el que se calculaba un valor de P.563 para cada locución (realizando un promedio de todas las tramas) , en este proyecto se dividirán las locuciones en fragmentos de 3 segundos y se calculará la P.563 para cada uno de los fragmentos. Más tarde, en el siguiente capítulo se explicará más detalladamente cómo se ha realizado el cálculo de esta medida para diferentes zonas de una locución.

3.3.2. SNR

Para la realización de este proyecto se utilizará la estimación de la SNR basada en los silencios de las locuciones. Este método hace uso de un detector de actividad (VAD) para identificar las tramas de voz y los silencios. Si queremos calcular la SNR media de una locución, se calcula la energía de cada una de las tramas y, atendiendo a la clasificación anterior, se calcula la energía media de las tramas de voz y la energía media de las tramas pertenecientes a silencios. De esta manera, se puede obtener la SNR como:

$$\text{SNR}=10\cdot\log\left(\frac{P_{voz}}{P_{silencios}}\right)$$

Donde P_{voz} y $P_{silencios}$ son las potencias correspondientes a las tramas de voz y silencio respectivamente. Posteriormente se describirá detalladamente cómo se ha calculado la SNR de las diferentes regiones de la locución.

3.3.3. KLPC

La *kurtosis* es una medida estadística que se usa, entre otras cosas, para calcular cuánto se aproxima cierta distribución a una gaussiana y, más específicamente, cómo de picuda es esta distribución. Se puede calcular utilizando el momento estadístico de cuarto orden de la distribución:

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

Donde a_p serían los valores cuya distribución se quiere evaluar, σ la desviación típica medida sobre la muestra y P el número de coeficientes LPC que se han obtenido de la trama (de una trama de 20ms se extraerán 21 coeficientes). La kurtosis que va a ser utilizada en este proyecto es la que se realiza sobre los coeficientes LPC, que se calcularán sobre las tramas de voz de cada locución. Como se describirá posteriormente, la KLPC se calculará sobre cada una de las tramas (cuya duración es de 20ms) de las locuciones y no globalmente, como hacían otros estudios previos.

4

Cálculo y estudio de las medidas de calidad

4.1. Cálculo de las medidas de calidad

Como se ha explicado anteriormente, el cálculo de las medidas de calidad no se va a realizar sobre cada locución de forma global sino que se dividirá dicha locución en tramas y se obtendrá un valor de cada indicador de degradación en cada una de ellas.

4.1.1. Funcionamiento del detector de actividad

Tanto la *SNR* como la *KLPC* hacen uso de un detector de actividad a la hora de realizar sus diferentes cálculos.

Este VAD divide la locución en fragmentos de 20 milisegundos y calcula la energía de cada uno de ellos. Una vez calculada la energía de las tramas se define un umbral de energía τ, definido por la siguiente fórmula:

$$\tau = E_{min.} + 0{,}4\,(E_{max.} - E_{min.})$$

La trama que no supere este umbral será reconocida como un **silencio** y la que sí lo supere será reconocida como una trama de **voz**.

4.1.2. División de las locuciones

Puesto que cada uno de los indicadores de degradación usados funciona de manera distinta, se ha realizado una división de las locuciones distinta para cada uno de ellos. En este apartado se detalla cada uno de ellos:

- **P.563:**

  El cálculo de la P.563 se realiza sobre tramas de 3 segundos y se procederá tal y como se describe describe a continuación:

  1. Se divide la locución en tramas de 3 segundos sin solapamiento.
  2. Se calcula la P.563 de cada trama, utilizando el software implementado a partir de la recomendación de la ITU, y se procederá a evaluar el resultado obtenido:
     - Si la trama está vacía o si, por el contrario, no está vacía pero los valores en la misma dan lugar a un mensaje de error por parte del medidor de calidad P.563 de la ITU, se le añaden otros 3 segundos y se vuelve a calcular la P.563. Si mediante este proceso de crecimiento de la trama, esta llega a alcanzar una duración mayor de 20 segundos o si resulta ser la última de la locución se le asigna el valor de la anterior trama etiquetada.
     - Si la trama no está vacía y tiene un valor numérico, se etiqueta la trama y se pasa a la siguiente.

  La siguiente figura (4.1) describe mediante un flujo de trabajo este proceso:

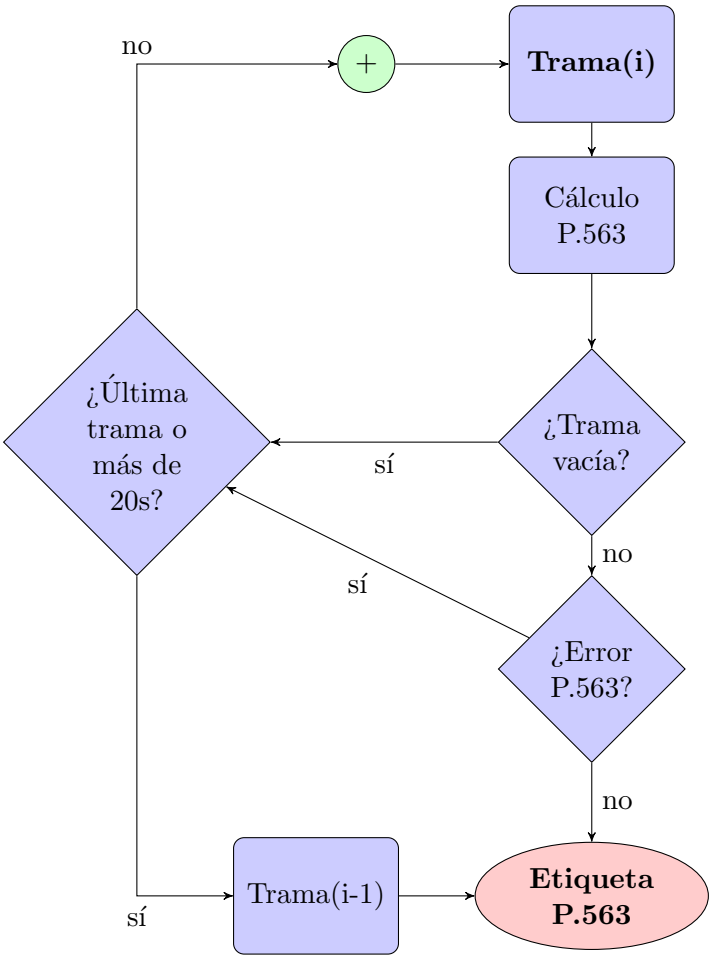

Figura 4.1: Esquema del cálculo de la P.563.

- **SNR:**

  Esta medida de calidad hace uso del VAD explicado anteriormente pues necesita conocer la energía de cada una de las tramas de voz y de los silencios. El cálculo se va a realizar sobre tramas de 3 segundos sin solapamiento de la siguiente manera:

  1. Detección de las tramas de voz y de los silencios (cada 20 milisegundos).
  2. Se divide la locución en tramas de 3 segundos (sin solapamiento).
  3. Se comprueba que la trama tiene un porcentaje de actividad de voz mayor o igual al 10.
     - Si el porcentaje de voz es mayor o igual del 10 % se pasa al siguiente punto.
     - Si el porcentaje de actividad es menor del 10 % se aumenta la trama en otras tres segundos y se vuelve a evaluar el porcentaje de actividad.
  4. Se calcula la SNR de la trama mediante esta fórmula, adaptada de la fórmula descrita en el apartado 3.3.2:

     $$\text{SNR}=10\cdot\log\left(\frac{E_{mediavoz}}{E_{mediasilencios}}\right)$$

     Donde $E_{mediavoz}$ y $E_{mediasilencios}$ se refieren a la energía media de voz y silencios, calculadas dividiendo la energía total de cada tipo de trama, obtenida tras la suma de cada una de las energías de cada trama, entre el número de tramas de cada tipo:

     $$E_{mediavoz} = \frac{E_{totalvoz}}{n^{o}tramasvoz}$$

La figura 4.2 muestra cómo se dividen las tramas de una locución para el cálculo de la SNR:

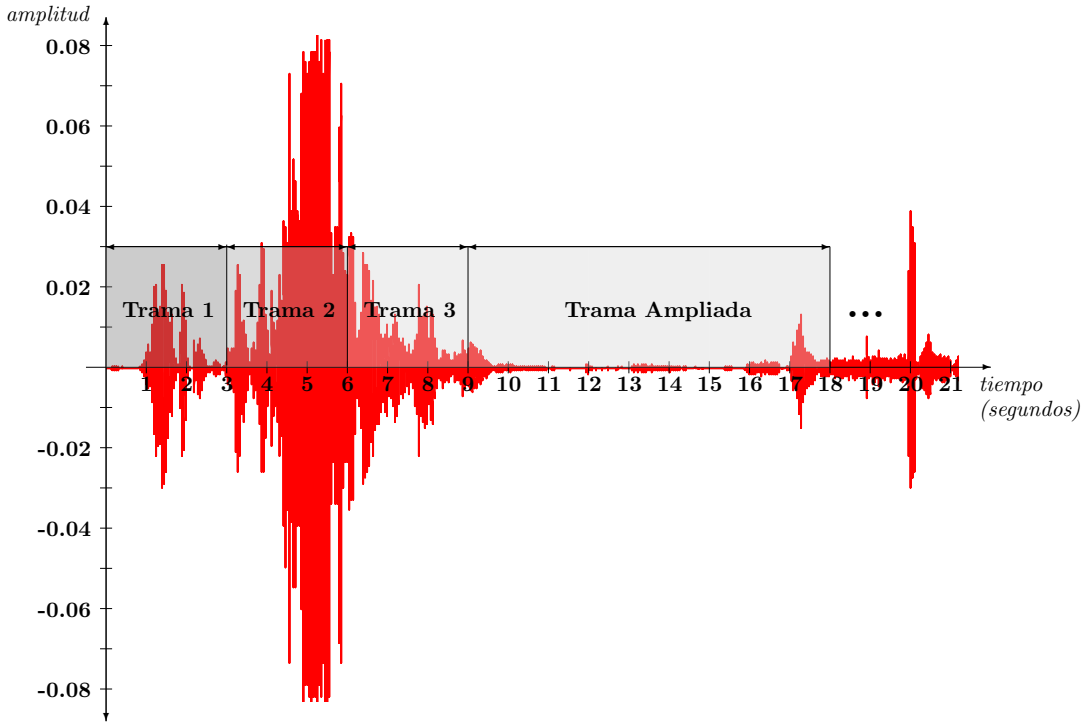

Figura 4.2: Ejemplo de la división en tramas de una locución (*SNR*).

- **KLPC:**

  La KLPC también necesita utilizar el VAD para diferenciar las tramas de voz y las de silencio, pues el cálculo de esta medida de calidad sólo se realizará sobre las tramas de voz. En este caso la forma de calcular el valor de cada trama será el siguiente:

  1. Detección de las tramas de voz y de los silencios (cada 20 milisegundos).
  2. Se divide la locución en tramas de 20 milisegundos sin solapamiento.
  3. Se calcula la KLPC sobre las tramas de voz. A las tramas que corresponden a silencios se les dará un valor de 3, que es el mínimo que puede llegar a alcanzar esta medida.

  La figura 4.3 muestra cómo se dividen las tramas de una locución para el cálculo de la KLPC:

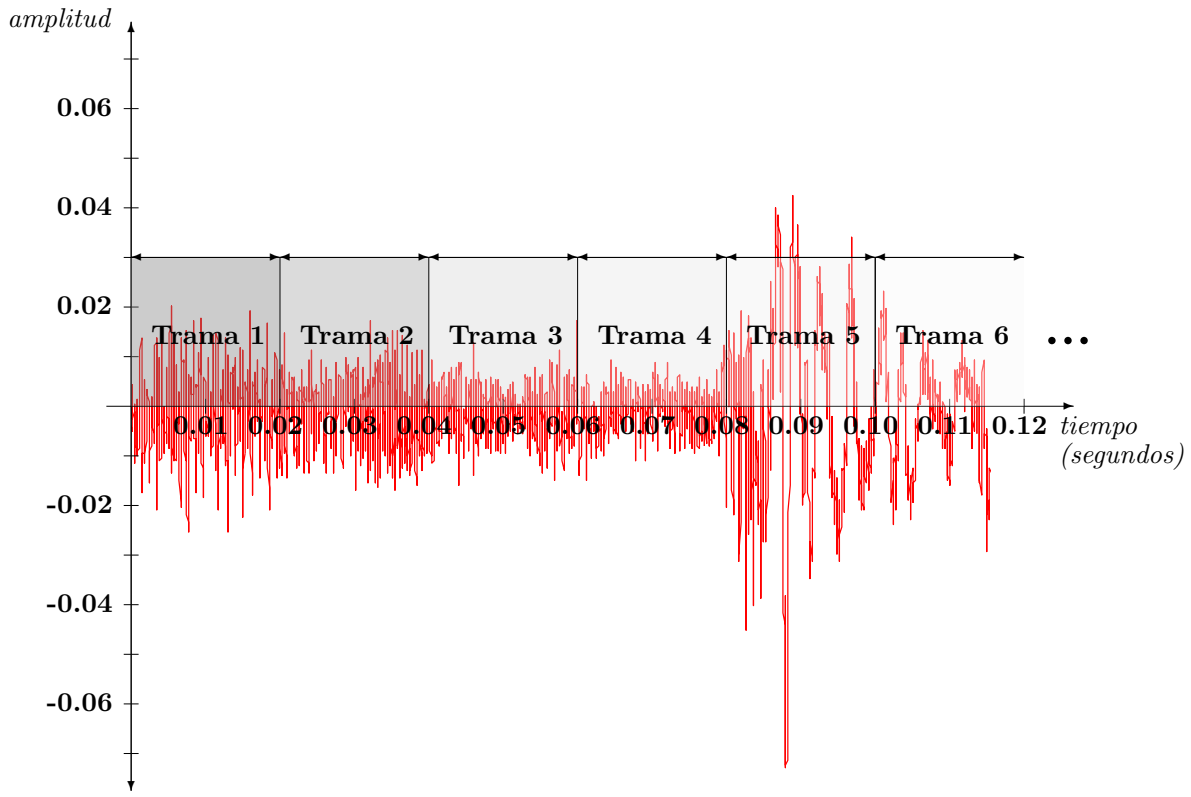

Figura 4.3: Ejemplo de la división en tramas de una locución (*KLPC*).

4.2. Transformación de las medidas de calidad

Tal y como se describe en [20], para poder estudiar estos indicadores como medidas de calidad propiamente dichas será necesario mapearlas para que su valor esté comprendido en un rango [0,1]. Para ello es necesario conocer el rango de valores en el que se encuentra cada una de las medidas originalmente. En la tabla 4.1 se puede observar el rango de valores entre los que varía cada una de las medidas tal y como se describen en [3]:

Medida	**Rango**
P.563	(1,5)
SNR	[0,60]
KLPC	[3,11]

Cuadro 4.1: Rango de valores antes del mapeo de las medidas de calidad.

Por tanto, para realizar el mapeo de la misma forma que se realiza en [3] se han utilizado las siguientes fórmulas:

$$Q_{SNR}(\mathrm{x}) = \frac{x}{60} \quad Q_{KLPC}(\mathrm{x}) = 1 - \left(\frac{x-3}{8}\right) \quad Q_{P563}(\mathrm{x}) = \frac{(x-1)}{4}$$

Una vez las medidas se encuentra en un rango normalizado es posible realizar una comparación entre ellas pues, cuanto más se acerque su valor a 1, mejor será la calidad de la muestra.

4.3. Combinación de las medidas de calidad

Para que el estudio de la calidad de una locución sea más sencillo es necesario contar con un único valor representativo de dicha calidad para así poder discriminar utilizando un único umbral. A la hora de combinar las distintas de calidad se ha decidido utilizar dos métodos: calculando la **media aritmética** y la **media geométrica**. Para realizar un estudio más exhaustivo se han calculado las medias de todas las combinaciones posibles entre las tres medidas utilizadas. En la tabla 4.2 se pueden observar todas las combinaciones usadas, así como la nomenclatura elegida para cada una:

	Media Aritmética	**Media Geométrica**
P.563 y SNR	Q_{MA}(P.563,SNR)	Q_{MG}(P.563,SNR)
P.563 y KLPC	Q_{MA}(P.563,KLPC)	Q_{MG}(P.563,KLPC)
SNR y KLPC	Q_{MA}(SNR,KLPC)	Q_{MG}(SNR,KLPC)
P.563, SNR y KLPC	Q_{MA}(SNR,P.563,KLPC)	Q_{MG}(SNR,P.563,KLPC)

Cuadro 4.2: Tabla con las distintas combinaciones entre las medidas de calidad

4.4. Experimentos de utilidad

4.4.1. Ejemplos ilustrativos del uso de las medidas por separado

Como una primera prueba se han evaluado los valores de las distintas medidas de calidad por separado para poder observar cómo se comportan frente a distintas locuciones. En todos los casos se han utilizado como locuciones de las pertenecientes a la base de datos **NIST2012** y elegidas aleatoriamente entre todas las disponibles.

Las siguientes gráficas representan de arriba hacia abajo: la onda de las locuciones seleccionadas, las tramas del VAD y las tramas de las tres medidas de calidad elegidas (todas ellas mapeadas). Se han usado tanto archivos telefónicos como archivos microfónicos para comprobar cual es el comportamiento de dichas medidas frente a los dos tipos de locuciones.

Locuciones telefónicas

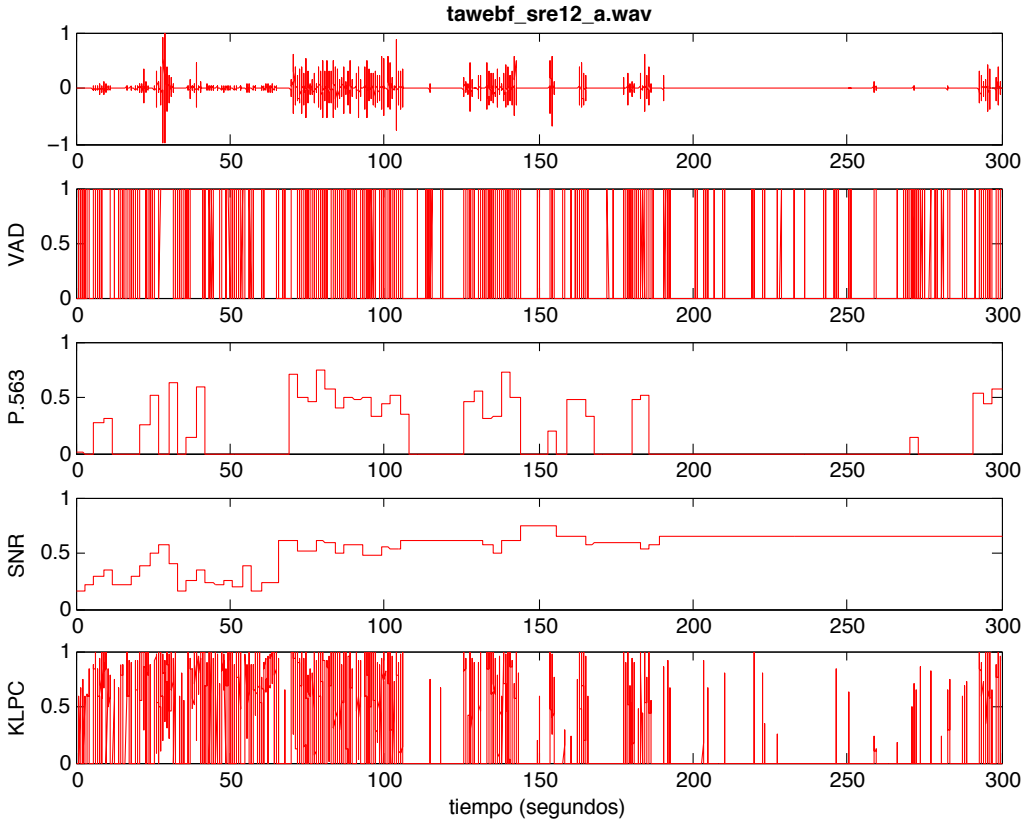

Figura 4.4: Gráficas representando las medidas de calidad de los archivos telefónicos elegidos.

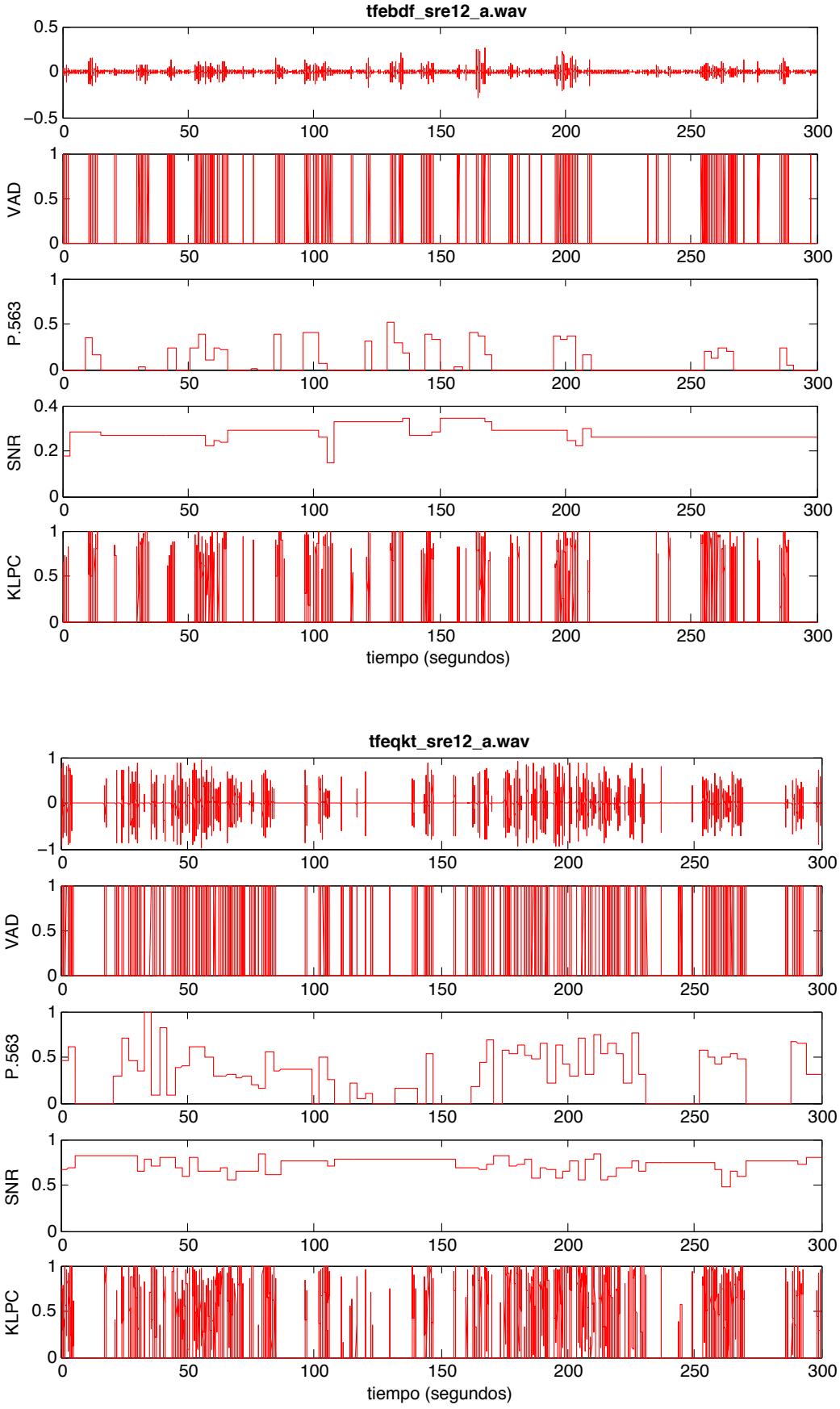

Figura 4.5: Gráficas representando las medidas de calidad de los archivos telefónicos elegidos.

Locuciones microfónicas

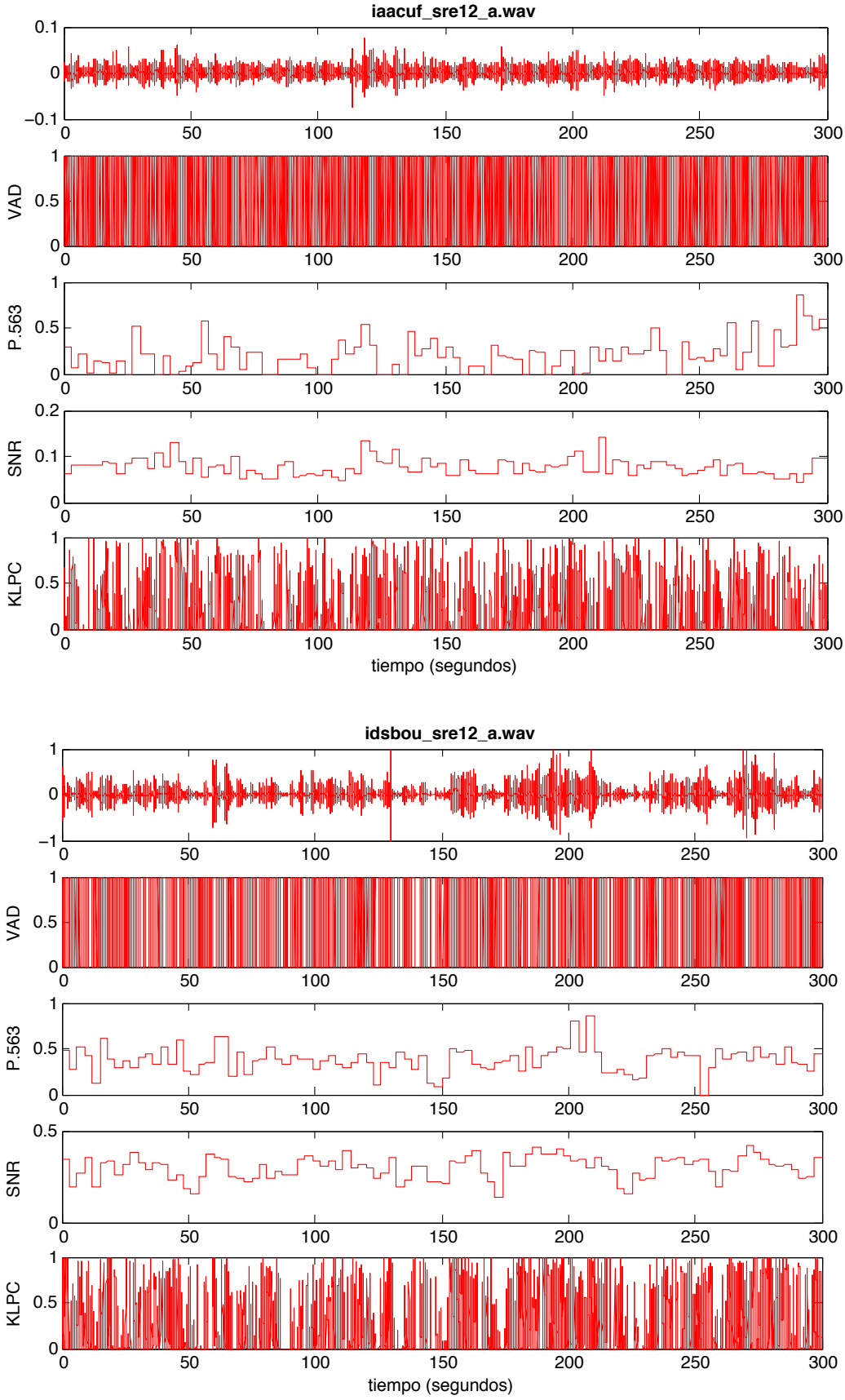

Figura 4.6: Gráficas representando las medidas de calidad de los archivos microfónicos elegidos.

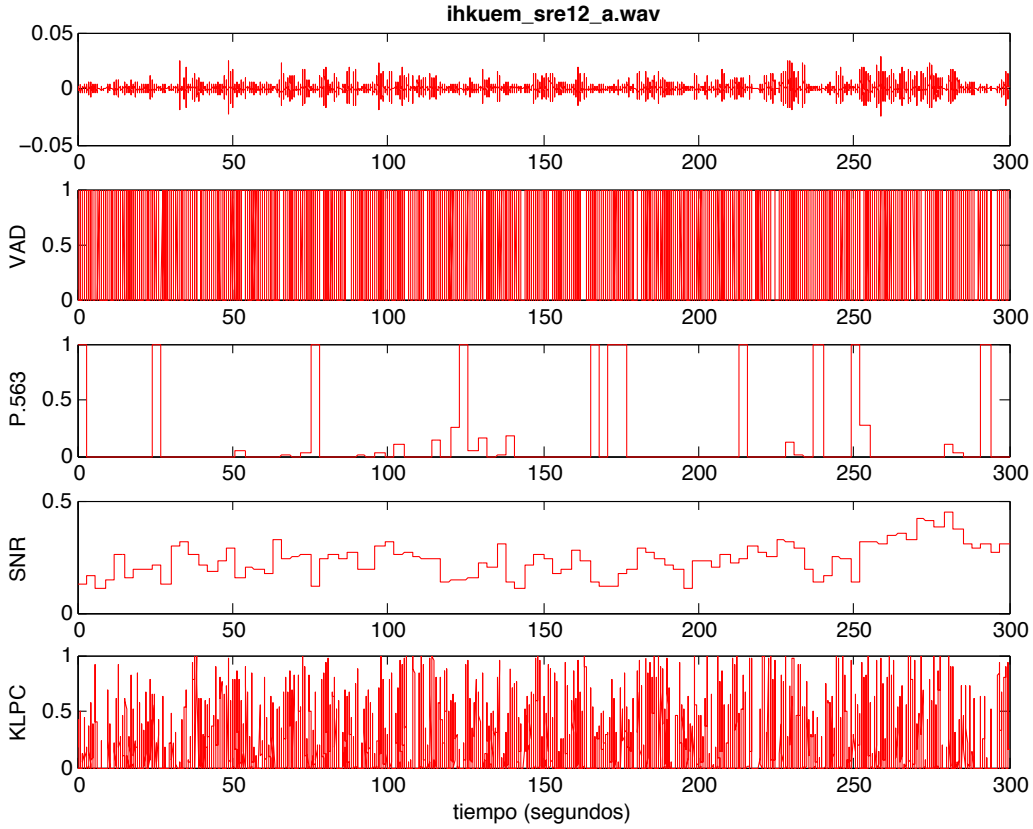

Figura 4.7: Gráficas representando las medidas de calidad de los archivos microfónicos elegidos.

En un primer estudio de las medidas de calidad por separado se puede observar que, en general, las locuciones elegidas son menos ruidosas que las microfónicas. Este ruido puede ocasionar que la SNR y la P.563 (esta última depende en buena parte de la SNR de la señal) tengan un peor comportamiento con este tipo de locuciones.

4.4.2. Evaluación de las medidas conjuntas

Una vez obtenida una única medida para cada trama de una locución, se puede estimar de manera empírica qué ficheros tienen peor calidad observando el valor de cada una de sus tramas. En la figura 4.2 se puede observar un posible histograma realizado sobre una locución en el que se representa el número de tramas frente al valor de la medida de calidad que se esté utilizando:

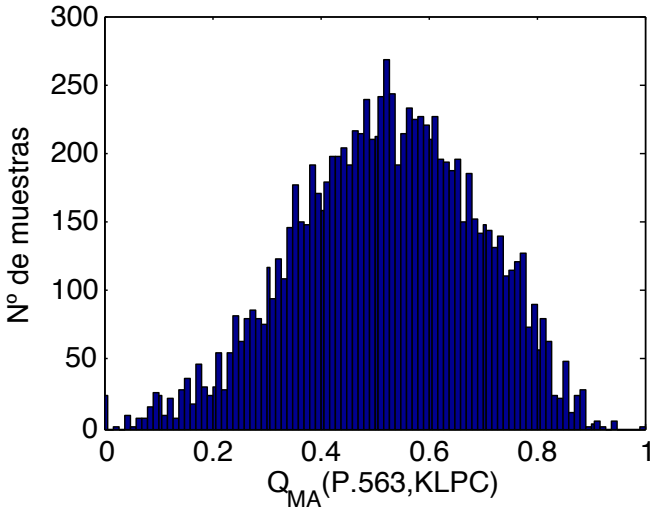

Figura 4.8: Ejemplo de histograma donde se representa el número de tramas frente al valor $Q_{MA}(P563, KLPC)$ de la locución tfeqkt_sre12_a.wav

En este caso, un segundo experimento fue el de realizar una estimación previa sobre la calidad *global* de ciertas locuciones evaluando el número de tramas que no llegaban a superar un valor de calidad conjunta de 0.2. Para poder realizar esta estimación primero fue necesario prescindir de las tramas que pertenecían a los silencios de la locución puesto que sólo interesaba evaluar la calidad de las tramas de voz. De esta manera se obtuvo un porcentaje de tramas de mala calidad como:

$$\textbf{\% de muestras de mala calidad} = 100 \cdot \frac{N_{<0,2}}{N_T}$$

Donde $N_{<0,2}$ se refiere al número de tramas cuyo valor no supera 0.2 y N_T son el número de muestras totales una vez han sido eliminados los silencios con el VAD. Para realizar esta prueba se han calculado las medidas de calidad conjuntas de varias locuciones tanto microfónicas como telefónicas, pertenecientes a la base de datos **NIST2012**.

Las siguientes figuras (figura 4.9 y 4.10) representan el número de locuciones en función del porcentaje de tramas de mala calidad para cada una de las medidas de calidad, para un número de locuciones telefónicas (6.226 locuciones) y microfónicas (9.905 locuciones) pertenecientes a la base de datos anteriormente citada.

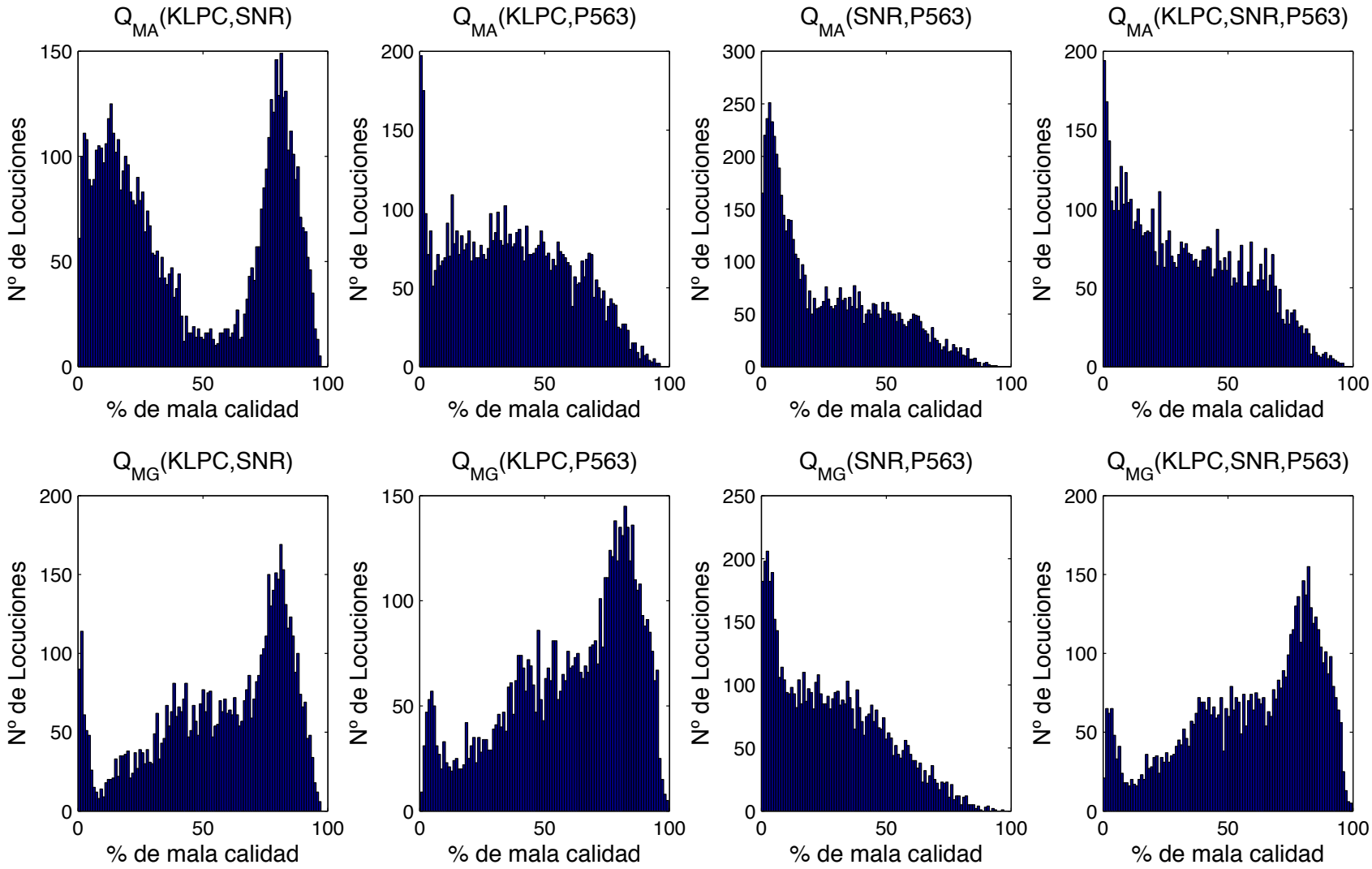

Figura 4.9: Histogramas que representan el número de locuciones telefónicas en función del porcentaje de tramas de mala calidad.

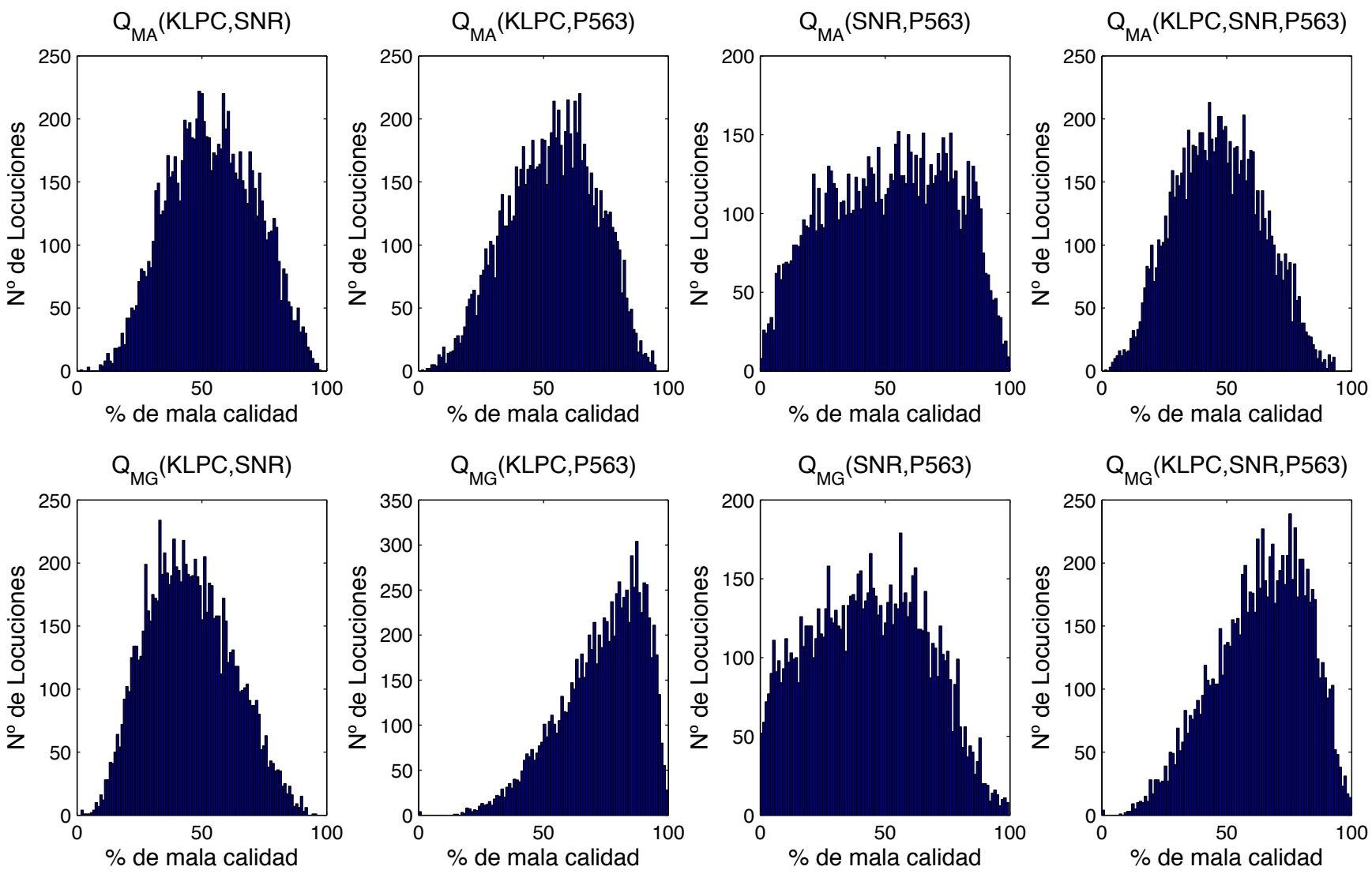

Figura 4.10: Histogramas que representan el número de locuciones microfónicas en función del porcentaje de tramas de mala calidad.

Se puede observar que las locuciones presentan, en el caso de las locuciones telefónicas, unos porcentajes de calidad de tramas bastante diferentes dependiendo de cuál sea el tipo de medida de calidad utilizada. Por ejemplo, las locuciones a las que se les ha aplicado la media aritmética presentan un mayo número de tramas por encima de 0.2 mientras que, en el caso de la media geométrica, el número de ficheros de mala calidad es, en general, bastante mayor.

En el caso de las locuciones microfónicas, los porcentajes de calidad de las muestras se encuentran, en la mayoría de los casos (sobre todo en las medidas que hacen uso de la media aritmética) en torno al 50 %. Según este estudio previo podría suponerse que, en la mayoría de las locuciones microfónicas, la mitad de las muestras de voz tienen un valor de Q por debajo de 0.2, de forma que, a la hora de decidir las muestras que se descartan antes de ser utilizadas en un sistema de reconocimiento de locutor, estaríamos desechando la mitad de las muestras de una locución.

En las siguientes figuras se pueden observar los histogramas para cada una de las medidas conjuntas y los valores para cada trama de las locuciones de prueba, ilustrando de forma un poco más particular el impacto que tendrá sobre las tramas de una locución el uso de dichas medidas de calidad.

Locuciones telefónicas: media aritmética

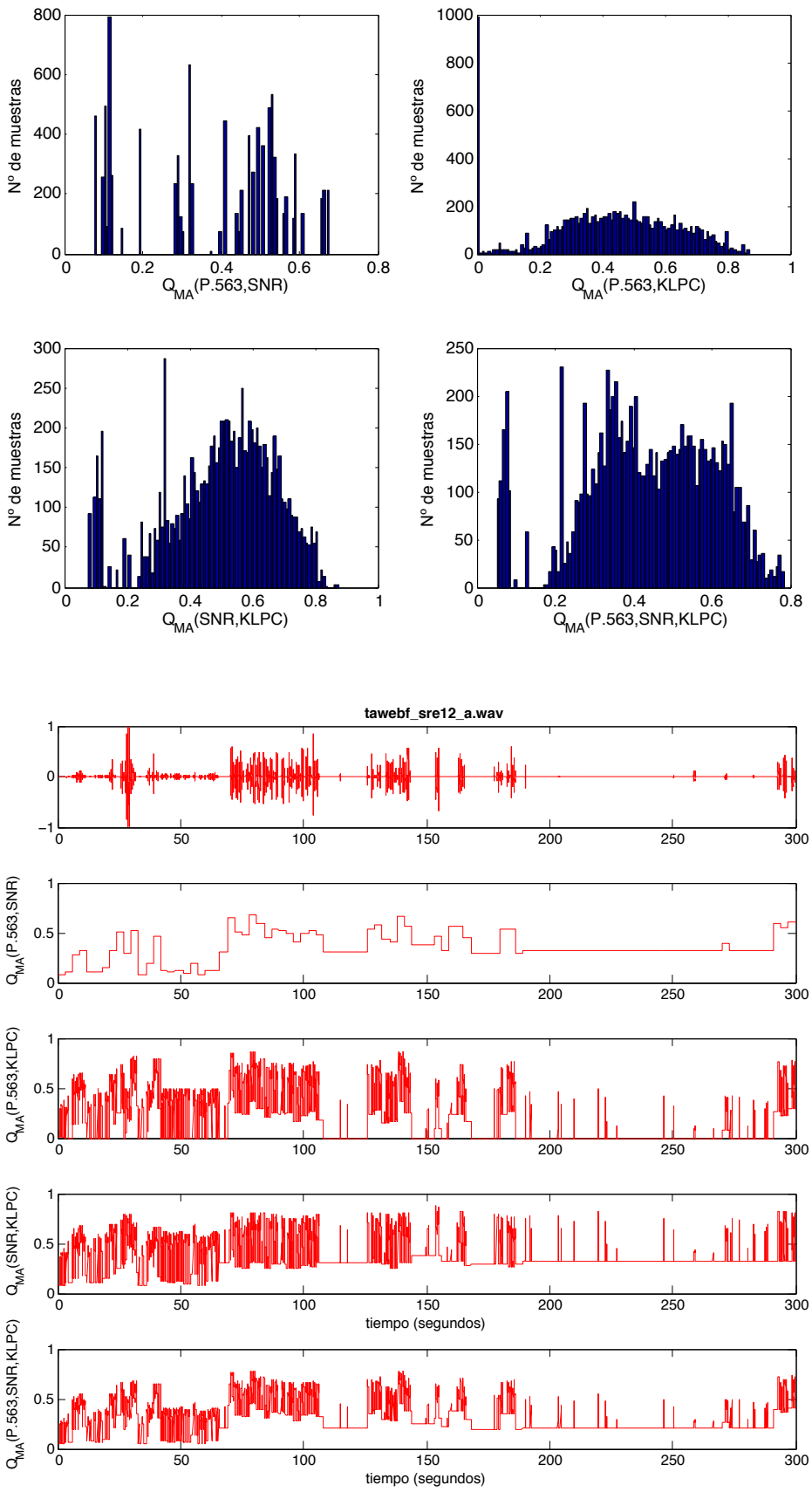

Figura 4.11: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

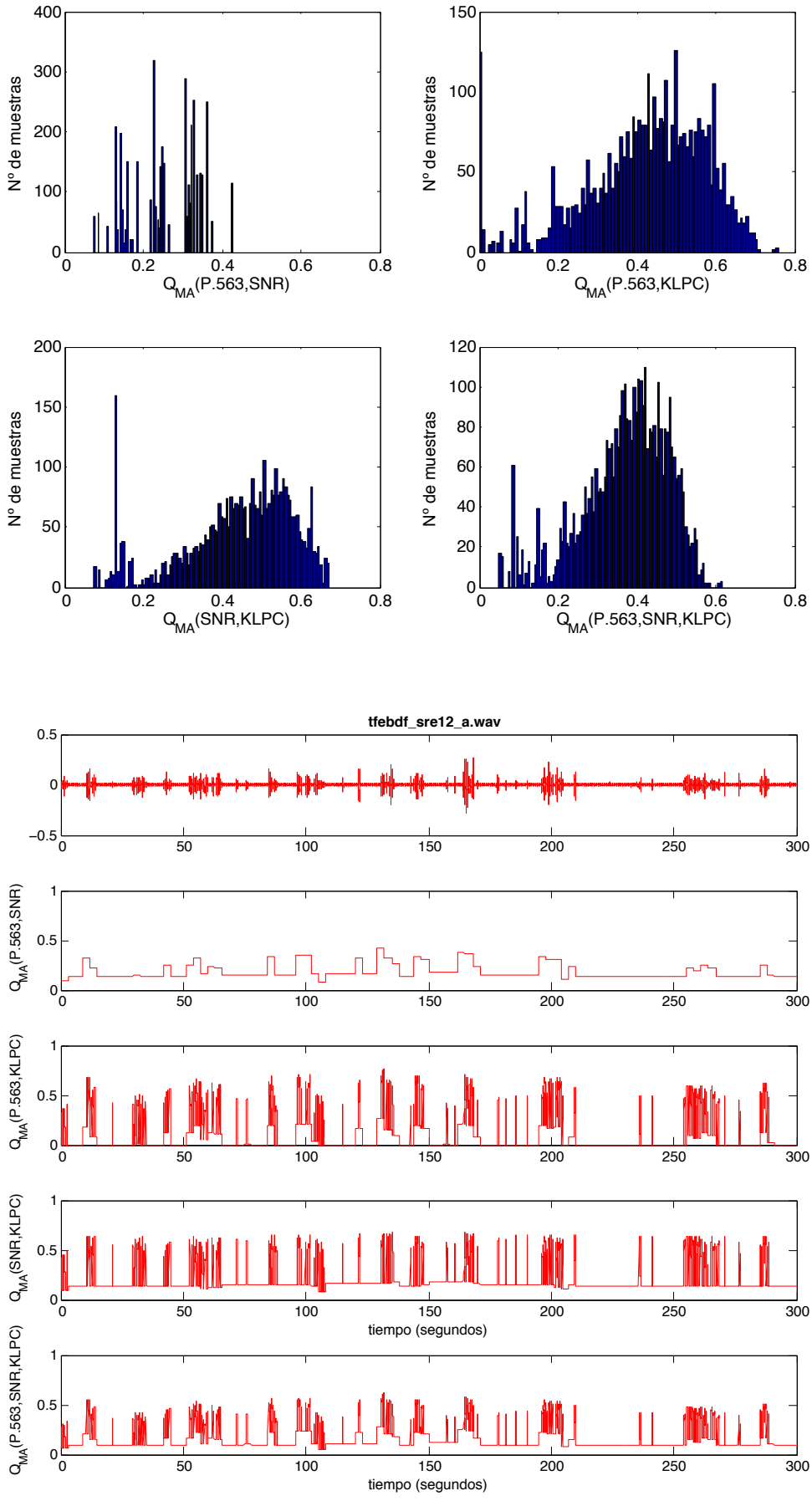

Figura 4.12: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

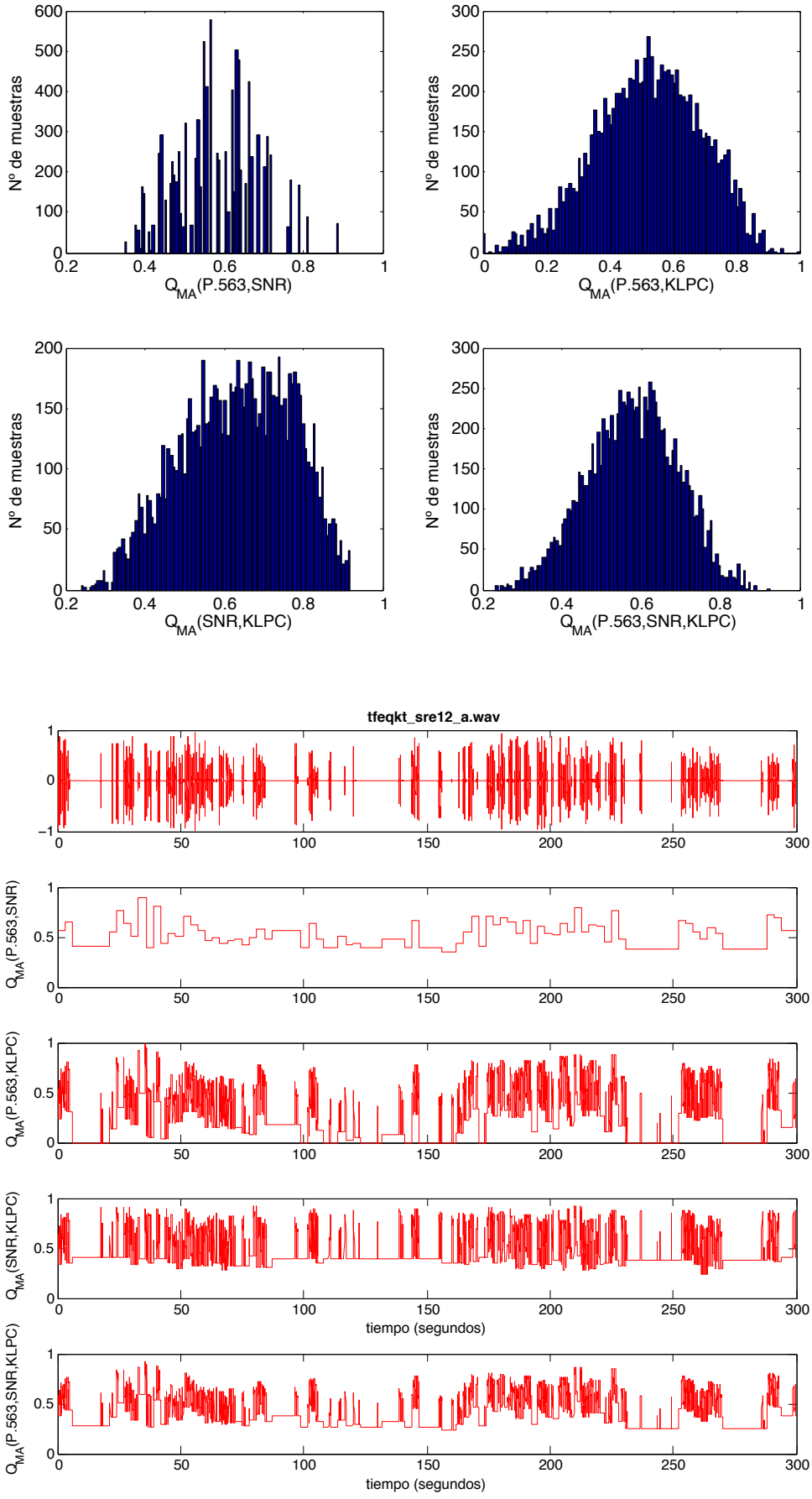

Figura 4.13: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

Media geométrica

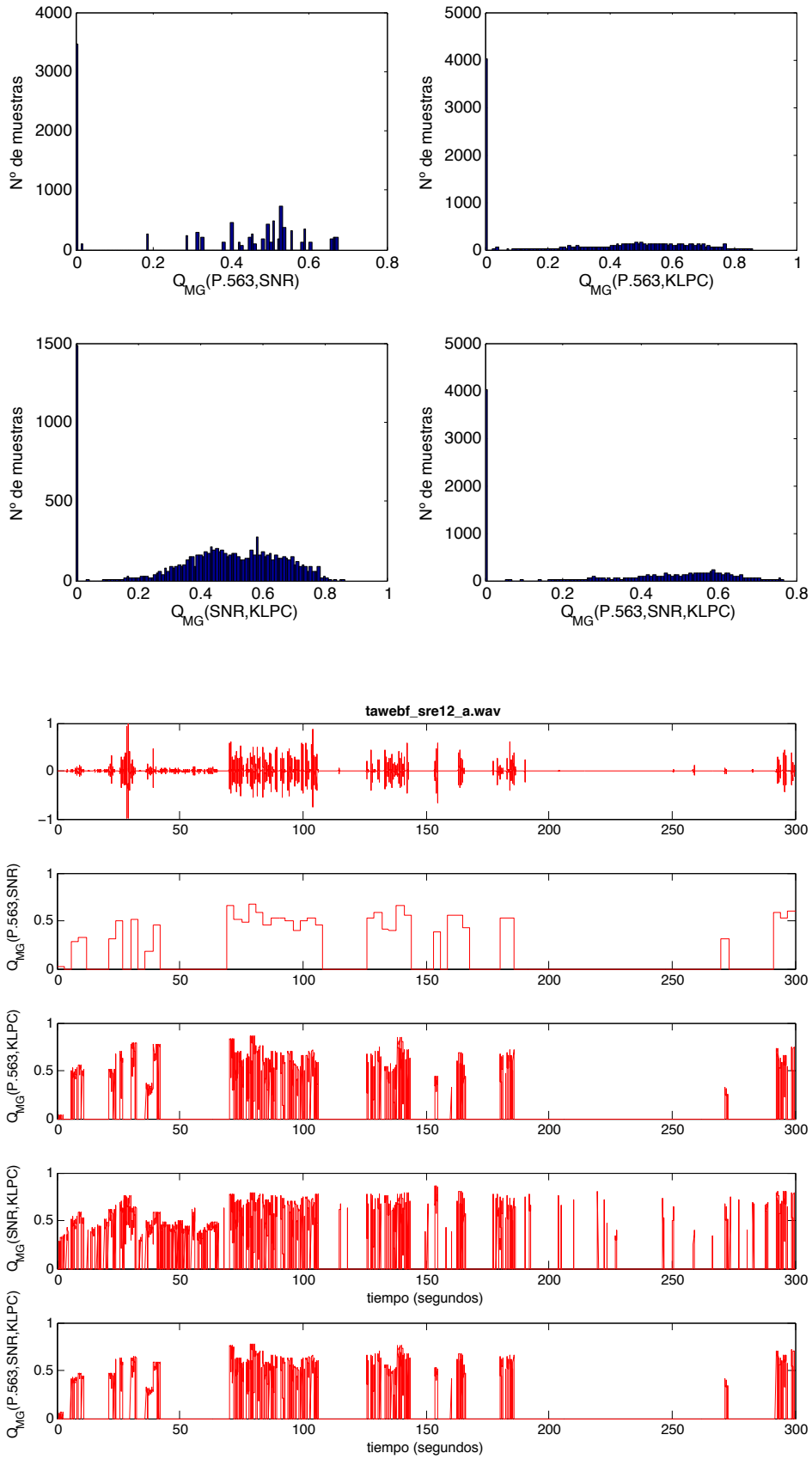

Figura 4.14: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

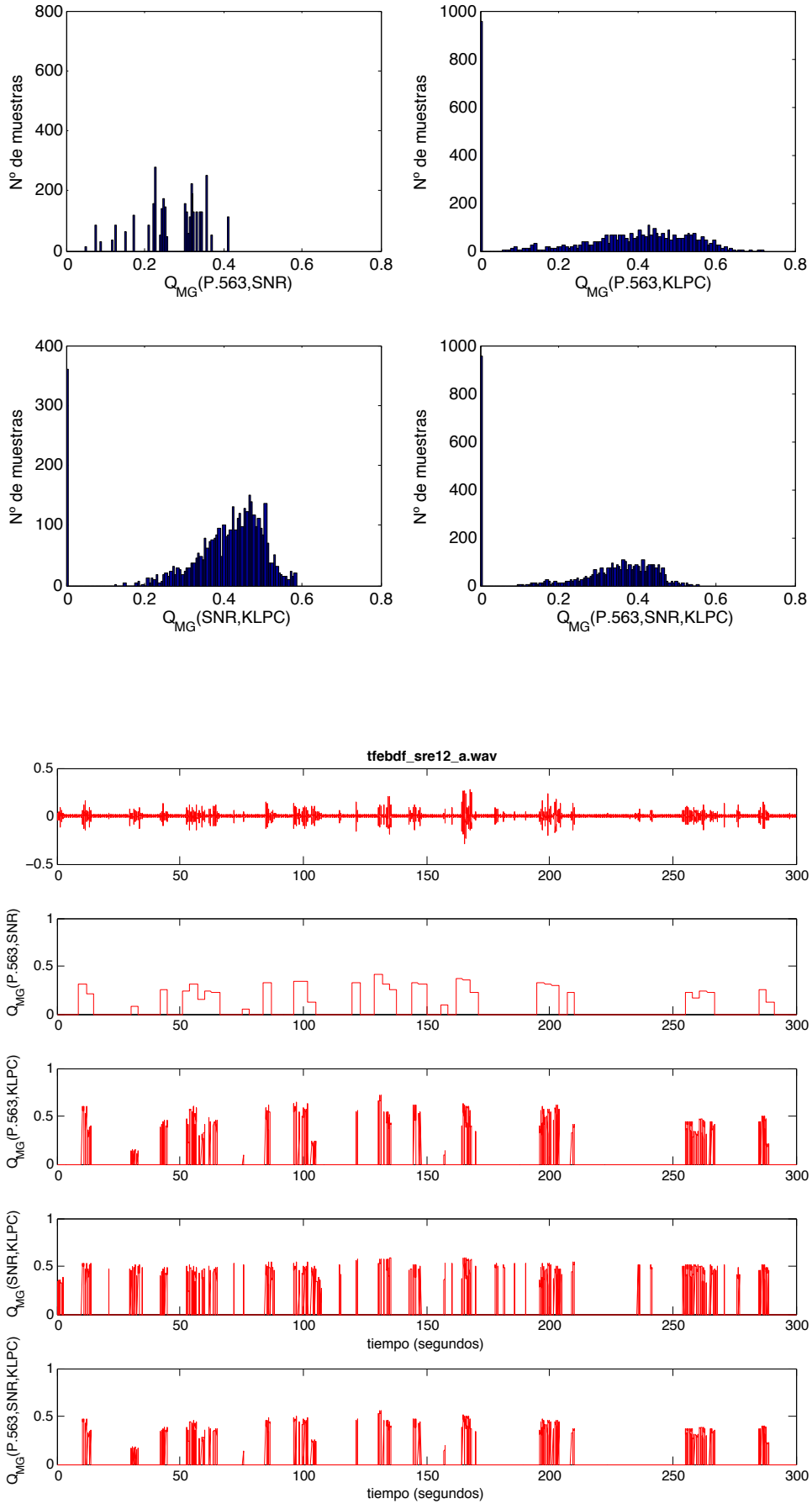

Figura 4.15: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

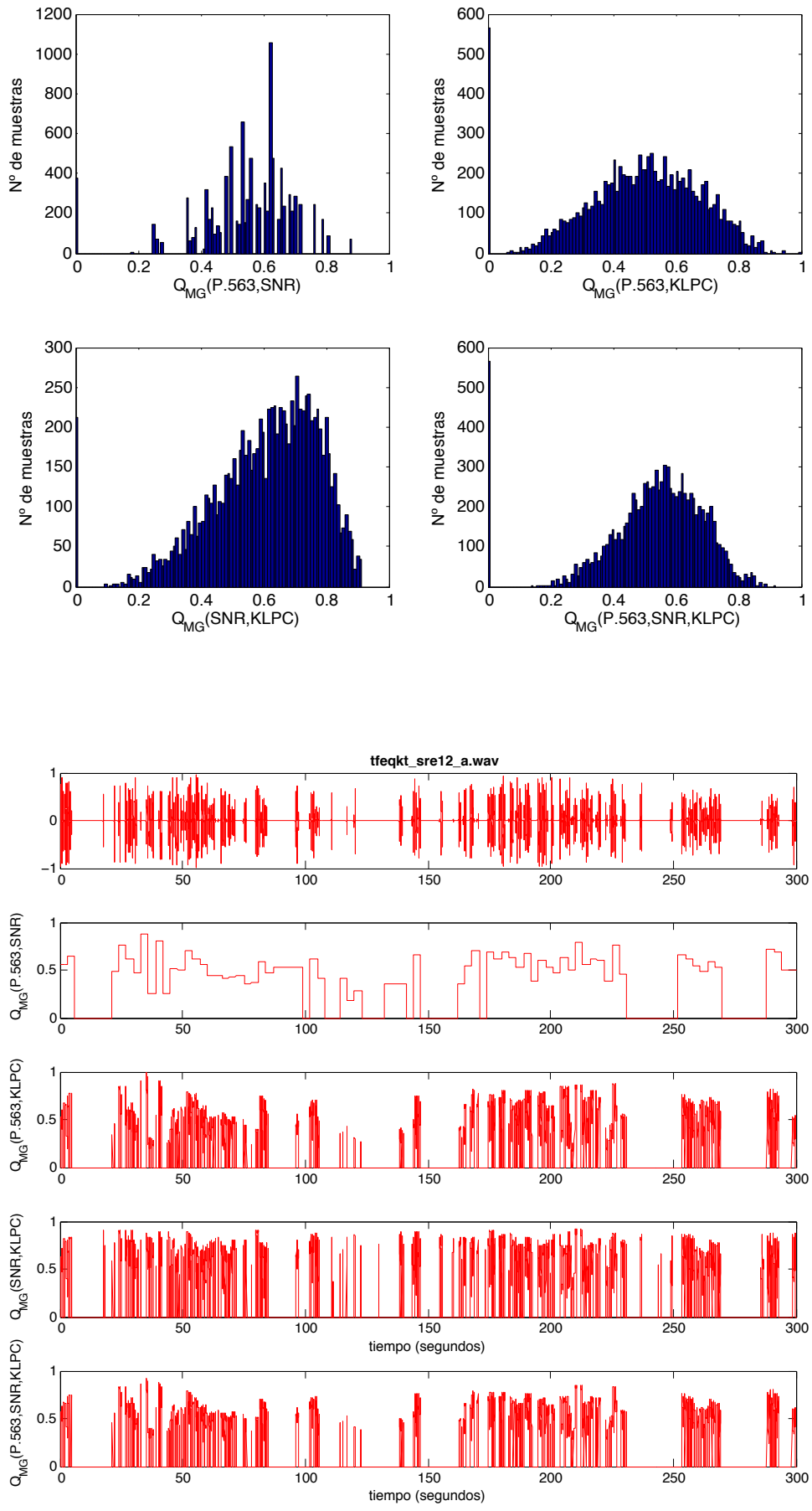

Figura 4.16: Gráficas e histogramas de las medidas de los archivos telefónicos elegidos.

Locuciones microfónicas: media aritmética

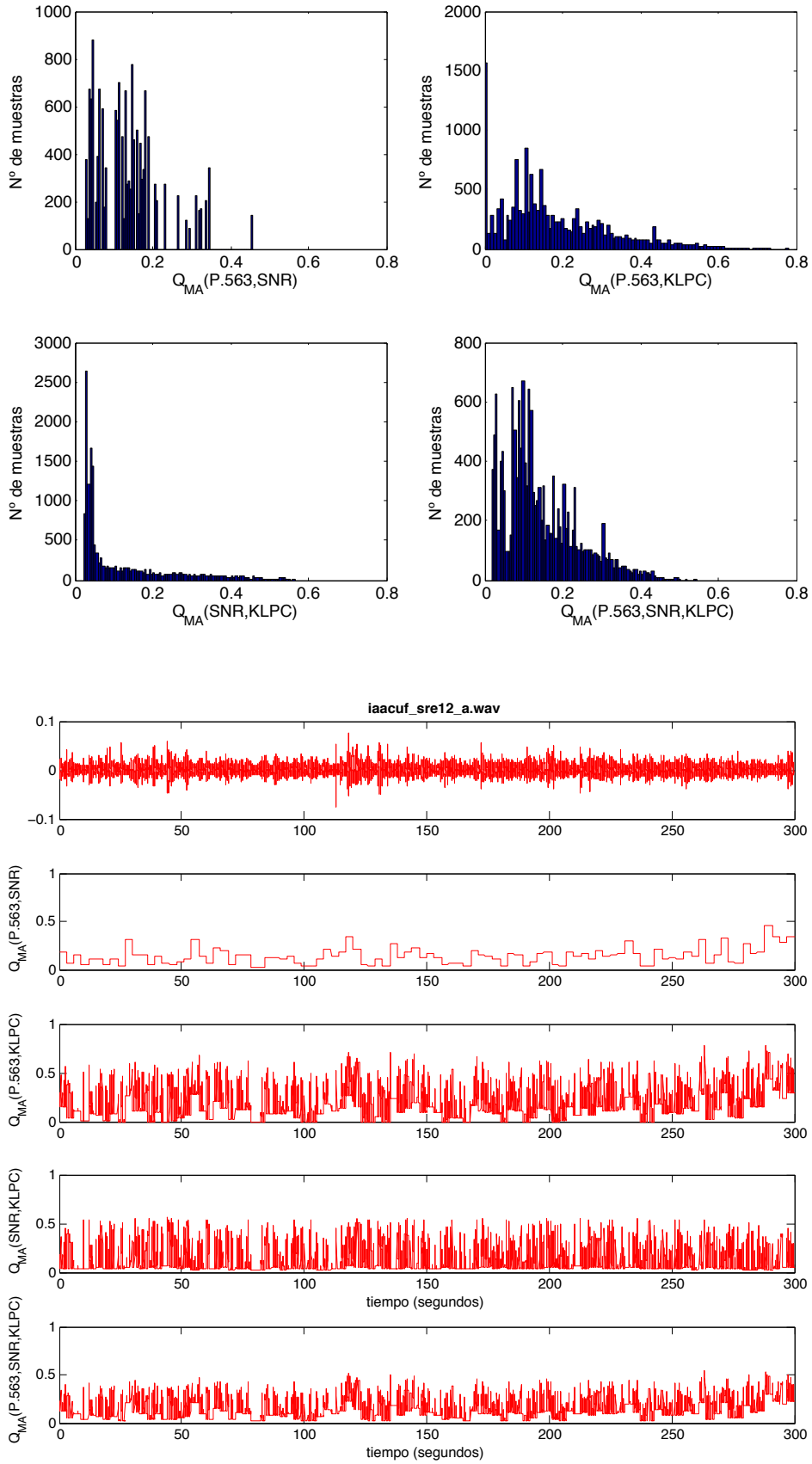

Figura 4.17: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

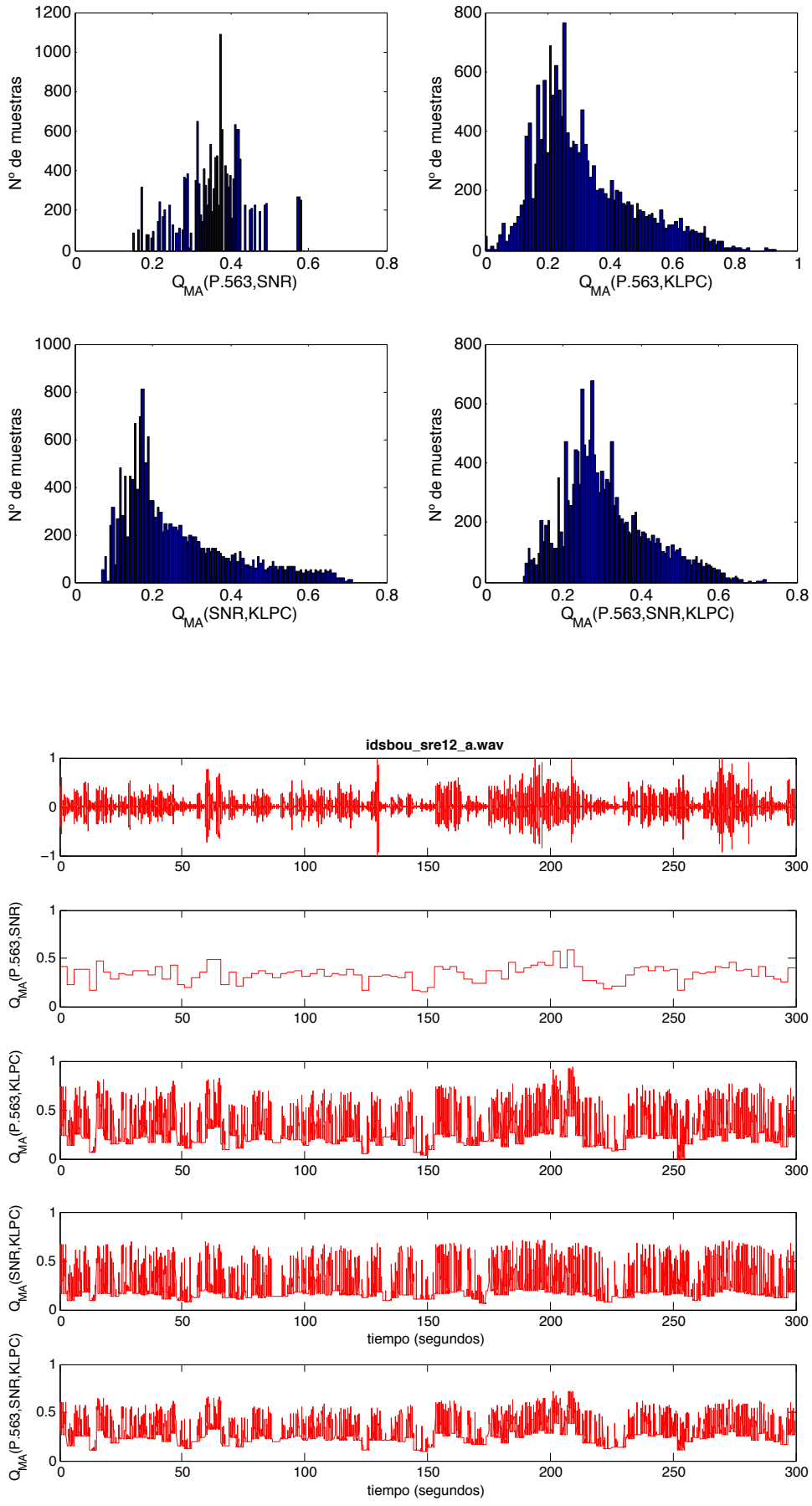

Figura 4.18: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

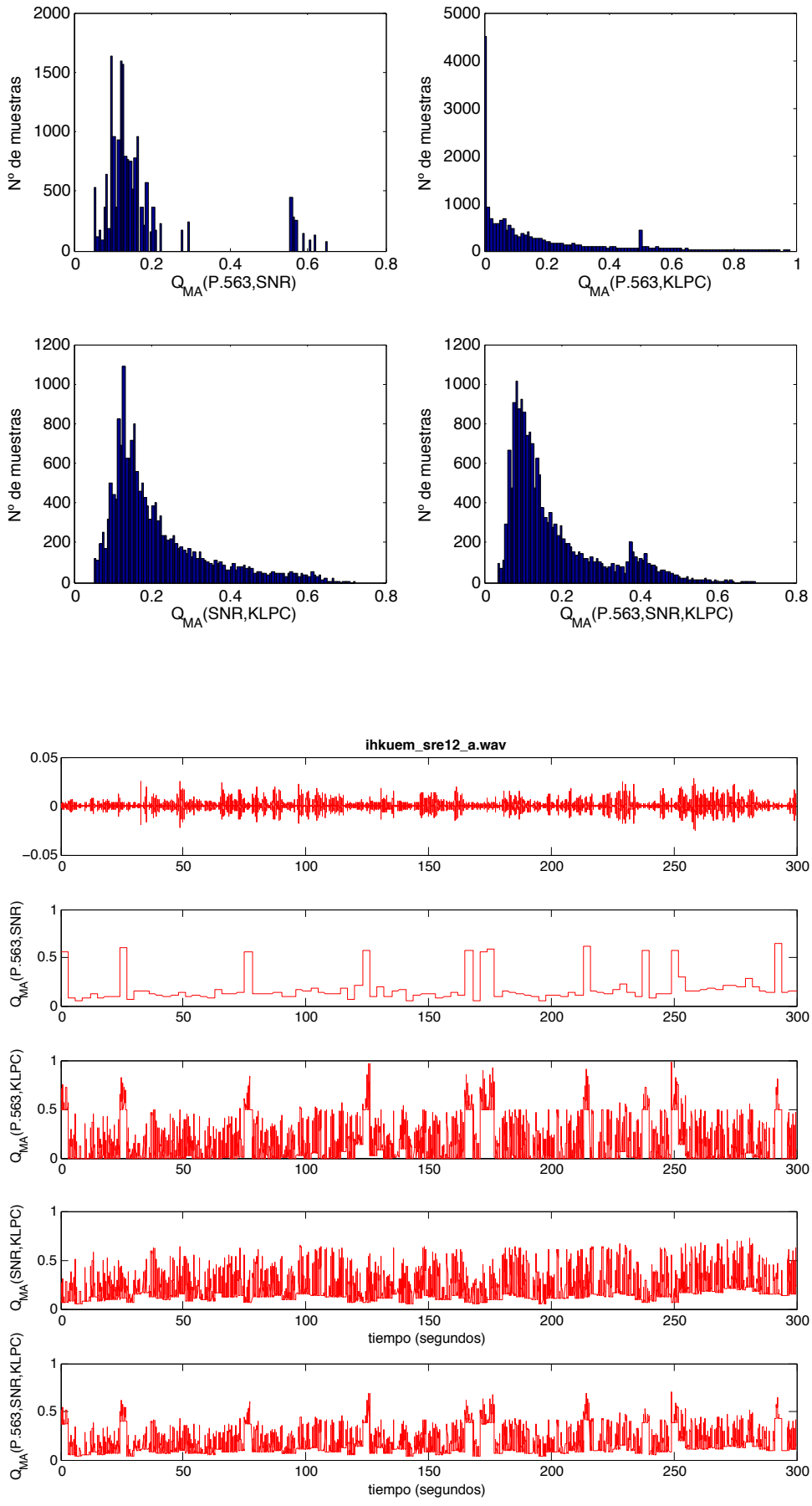

Figura 4.19: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

Media geométrica

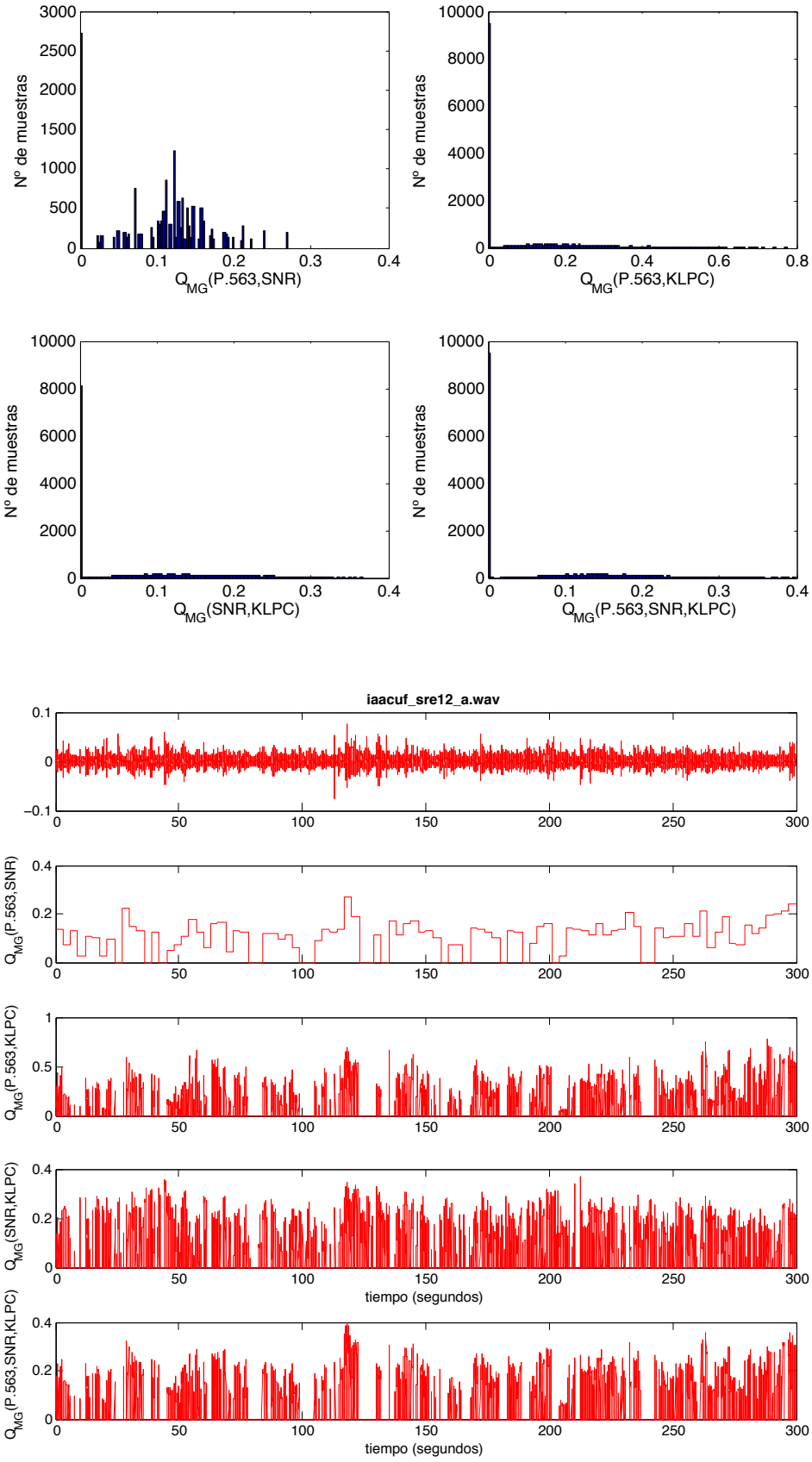

Figura 4.20: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

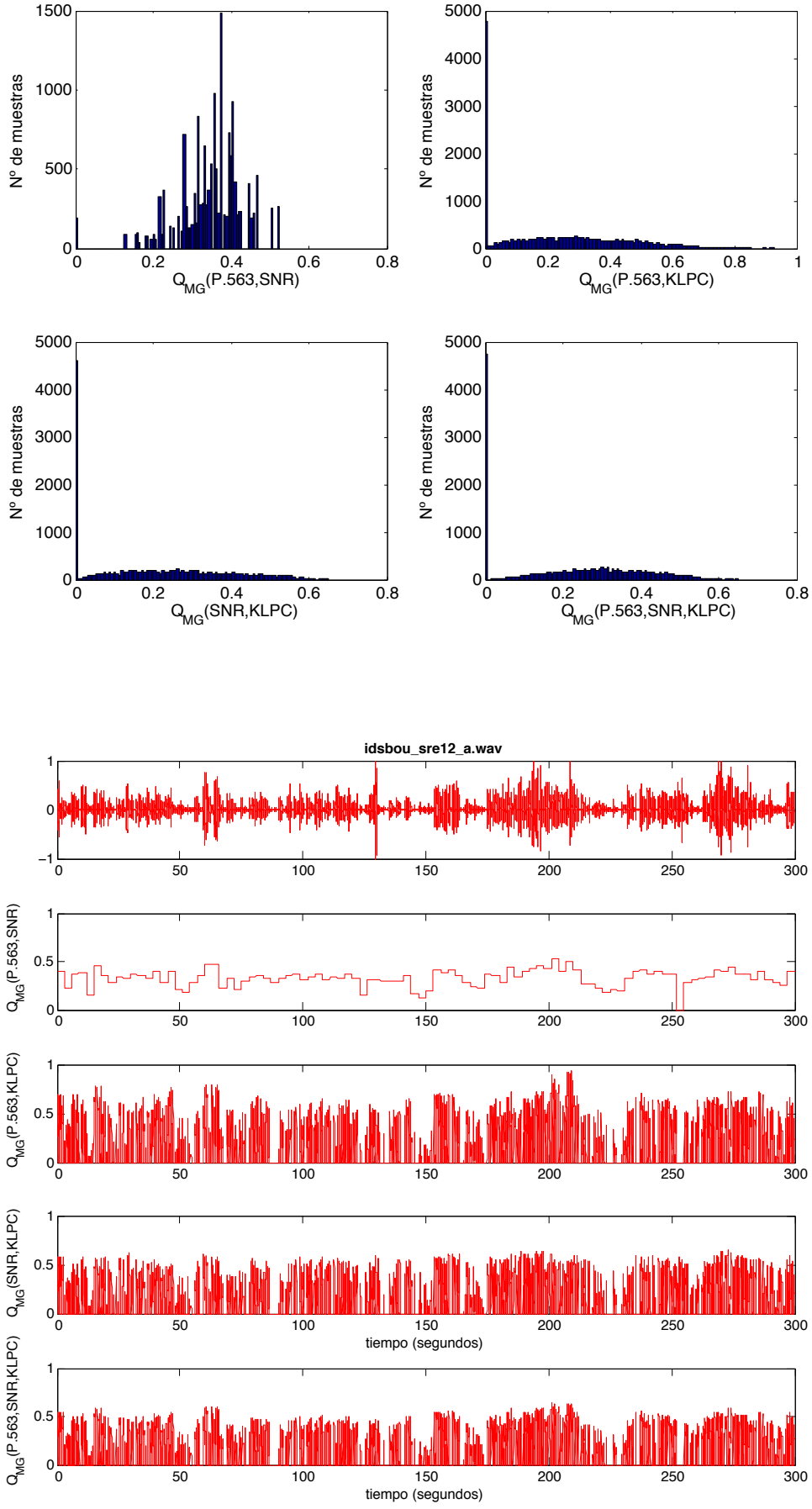

Figura 4.21: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

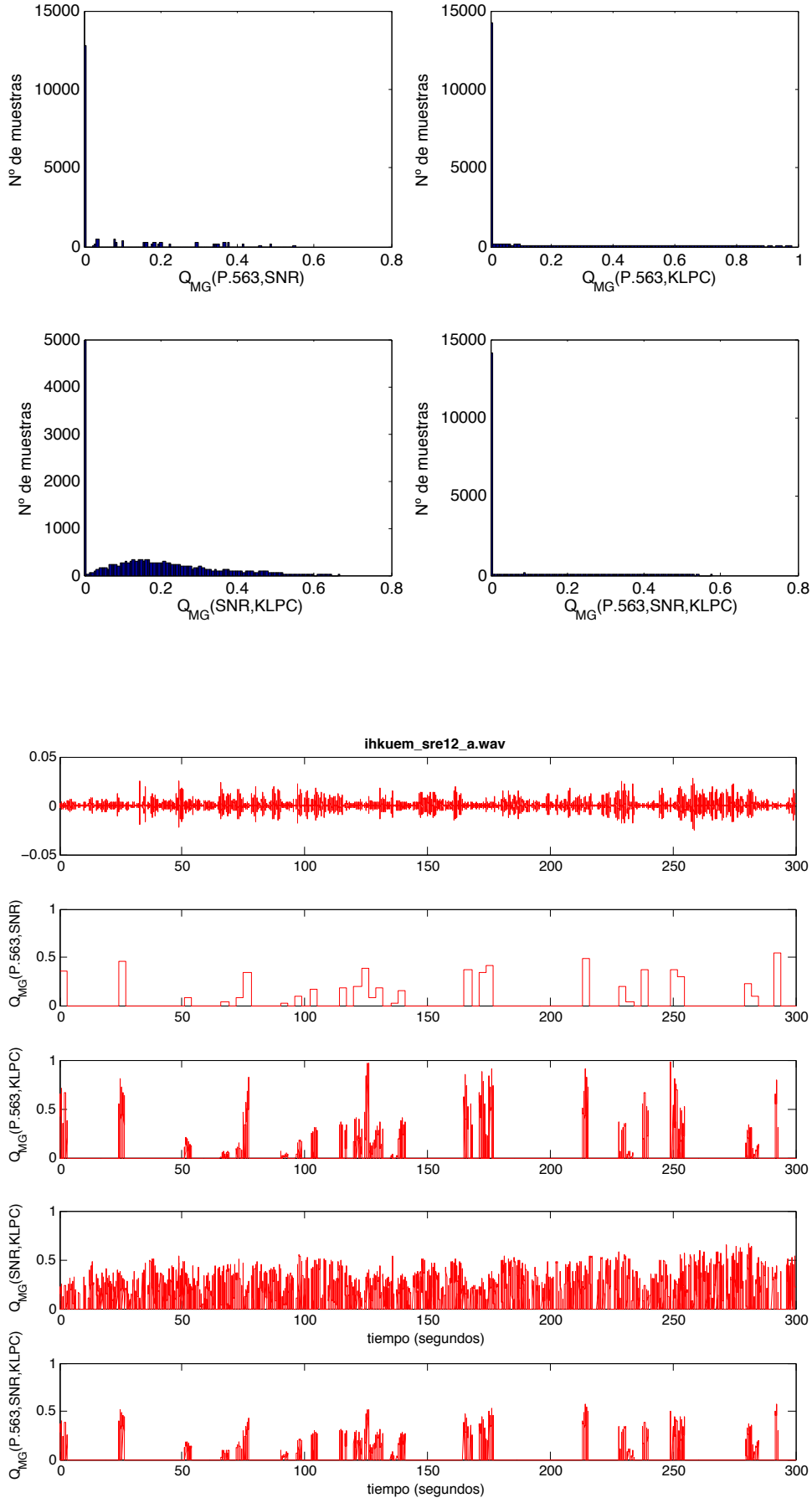

Figura 4.22: Gráficas e histogramas de las medidas de los archivos microfónicos elegidos.

En los archivos utilizados, se ha observado que en general las medidas de calidad que utilizan la media aritmética son menos restrictivas que las que hacen uso de la media geométrica, dando lugar a un mayor número de tramas por encima de 0.2. Además, las locuciones microfónicas cuentan con un mayor número de tramas por debajo de 0.2 y es posible que den peores resultados durante los experimentos que las locuciones telefónicas.

5

Experimentos Realizados y Resultados

5.1. Bases de datos y protocolo

La base de datos y protocolos **NIST SRE 2012** [21], 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 test como para entrenamiento de modelos.

Las condiciones de test y entrenamiento de modelos incluyen conversaciones grabadas sobre un canal telefónico, y sobre un canal microfónico en el escenario conocido como entrevista (int), en el cual existe el locutor principal y un entrevistador que formula preguntas, y adicionalmente locuciones de test grabadas sobre un canal microfónico.

Dentro del protocolo de evaluación se definen cuatro condiciones: tlf-tlf, mic-mic, tlf-mic, y mic-tlf. En los experimentos de este proyecto se van a utilizar dos de las cuatro condiciones anteriormente citadas:

- tlf-tlf
- mic-mic

El número de enfrentamientos para cada condición utilizada en este proyecto se puede observar en la siguiente tabla:

Canal entrenamiento	Canal test	Número de enfrentamientos	Target	Non-Target
Telefónico	Telefónico	205.639	4.258	201.381
Microfónico	Microfónico	22.260	1.041	21.219

Cuadro 5.1: Número de enfrentamientos por tipo de canal y según Target y Non-Target

El número de ficheros de test y train que se van a utilizar, así como la duración de cada uno, se pueden observar en las siguientes tablas:

	Número de ficheros	Duración
Test	6.226	300 segundos
Train	7.207	10, 100 y 150 segundos

Cuadro 5.2: Número y duración de las locuciones telefónicas de test y train

	Número de ficheros	Duración
Test	9.905	300 segundos
Train	5.536	180 y 480 segundos

Cuadro 5.3: Número y duración de las locuciones microfónicas de test y train

5.2. Experimentos realizados

La finalidad de los experimentos que se han realizado en este proyecto es la de evaluar el rendimiento del sistema desarrollado sobre un sistema de reconocimiento de locutor utilizando los dos tipos de locuciones explicados en el apartado 5.1.

Para la realización de dichas pruebas se van a utilizar todas las medidas de calidad por separado y, además, todas sus combinaciones mediante la media aritmética y la geométrica. Todas estas medidas se van a combinar con las etiquetas de actividad de voz utilizadas durante la evaluación del NIST2012 que realizó el grupo ATVS. El sistema, a grandes rasgos, se puede contemplar en el siguiente diagrama de bloques (figura 5.1):

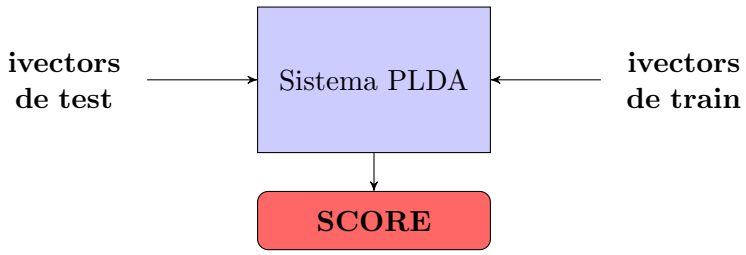

Figura 5.1: Esquema de alto nivel de la entrada de un sistema PLDA - *Probabilistic Linear Discriminant Analysis*.

Como las etiquetas del VAD están formadas por cadenas de 2 y 0, para poder unir estos ficheros con los correspondientes a las medidas de calidad de la locución, será necesario convertir estos últimos en otros con ese mismo formato. En este punto se hace necesario el uso de un **umbral de decisión**. Gracias a este umbral, se decidirá si una trama es adecuada o no (si tiene calidad suficiente o carece de ella) para pasar a formar parte del proceso de reconocimiento de locutor. La primera tanda de experimentos consistirá en variar este umbral, para cada una de las medidas de calidad, entre 0.1 y 0.9 para después elegir el umbral de valores que presente mejores resultados. Una vez que las etiquetas presentan el mismo formato es posible combinarlas mediante una operación lógica binaria de tipo **AND**.

Donde realmente va a entrar en juego el sistema desarrollado será durante el cálculo de los **ivectors** que se usarán para calcular la puntuación y el EER del sistema de reconocimiento de locutor. Además, otra de las características del sistema que se va a utilizar, y que ha sido diseñado por el ATVS, es que hace uso de un **filtrado Wiener** antes de realizar el cálculo de los ivectors para eliminar parte del ruido de las locuciones. El sistema desarrollado en este proyecto calculará las medidas de calidad de las locuciones antes de que estas pasen por el filtrado Wiener para así evaluar la locución sin haber sido modificada de ninguna manera. El sistema completo de cálculo de los ivectors se puede observar en la figura 5.2:

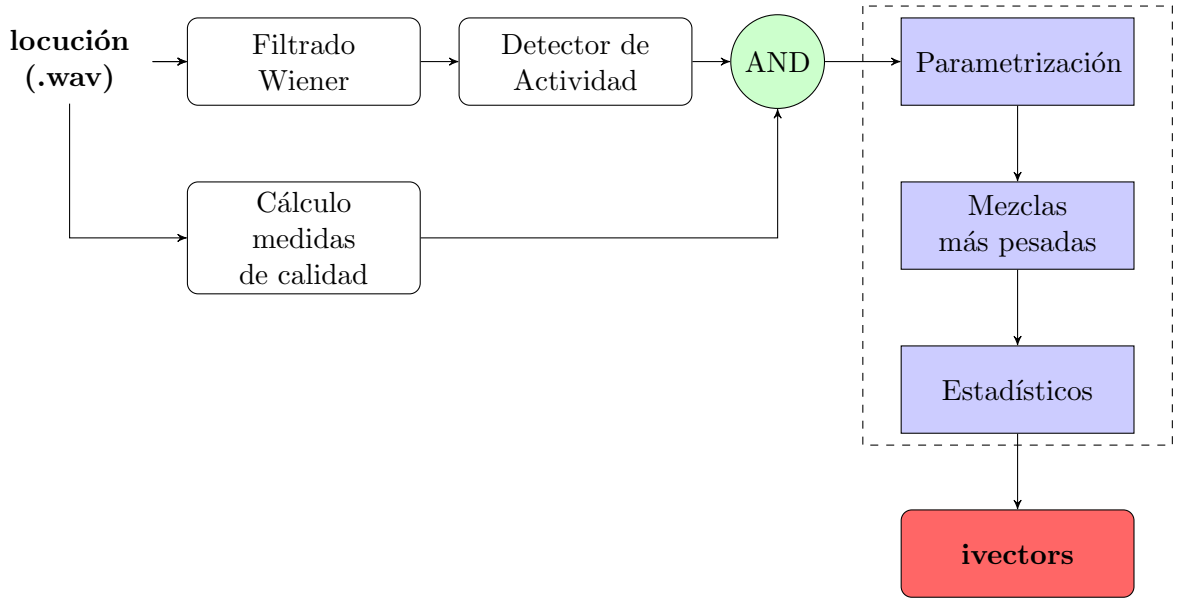

Figura 5.2: Esquema representando el cálculo de los ivectors.

A la hora de evaluar el rendimiento se calcularán las medidas de calidad de una serie de ficheros telefónicos y microfónicos y, una vez obtenidas las etiquetas conjuntas, se calculará la EER - *Equal Error Rate* del sistema para poder compararla con la EER cuando se usan únicamente las etiquetas del VAD para comprobar si ha habido alguna mejoría.

5.2.1. Estructura de los experimentos

De esta manera, una vez que se han calculado las medidas de calidad de las locuciones, los experimentos que se han realizado son los siguientes:

1. **Archivos telefónicos:**
   - Cálculo de las medidas de calidad de los ficheros de **test** y combinación con la etiquetas del VAD. A la hora de combinar las etiquetas, se variará el **umbral de decisión** entre 0.1 y 0.9 en intervalos de 0.1.
   - Comparación de los distintos valores de EER y elección de los valores de umbral que dan mejores resultados para poder así acotar las pruebas.
   - Cálculo de las mismas medidas de calidad con los ficheros de **train** del modelo PLDA para comprobar si la mejoría aumenta. El umbral que se ha usado en estas pruebas es el umbral acotado que se ha elegido durante las pruebas con los archivos de test.
   - Comparación de los nuevos valores de EER y elección de las medidas de calidad con mejores resultados.

2. **Archivos microfónicos:**

   - Cálculo de las medidas de calidad de los ficheros de **test** y combinación con la etiquetas del VAD.
   - Comparación de los distintos valores de EER y elección de los valores de umbral que dan mejores resultados para poder así acotar las pruebas.
   - Cálculo de las mismas medidas de calidad con los ficheros de **train** del modelo PLDA para comprobar si la mejoría aumenta.
   - Comparación de los nuevos valores de EER y elección de las medidas de calidad con mejores resultados.

5.3. Resultados obtenidos

5.3.1. Medidas de calidad en locuciones telefónicas

En este apretado se muestran los resultados obtenidos tras el cálculo de las medidas de calidad sobre la base de datos telefónicos **NIST SRE 2012**.

Resultados obtenidos usando locuciones de test

El primer experimento que se ha realizado ha sido el del cálculo de las medidas de las locuciones de test. Una vez calculadas las medidas de calidad de dichas locuciones y combinadas con los ficheros del VAD, se utilizarán dichos ficheros combinados a la entrada del sistema PLDA junto con los ivectors de train originales (sin utilizar las medidas de calidad) para comprobar si mejora o no el rendimiento del sistema.

Para poder comprobar esta mejora se comparará el EER obtenido tras las pruebas con las medidas de calidad con el EER original obtenido sin usar ningún tipo de medida de calidad. Este valor es:

$$EER_{SoloVAD}\text{=}\mathbf{5.9117}$$

Una vez obtenidos todos los valores de EER, se pueden dibujar las gráficas que comparan cada valor de EER por cada valor de la medida de calidad indicada. Un ejemplo de este tipo de gráfica sería la figura 5.3:

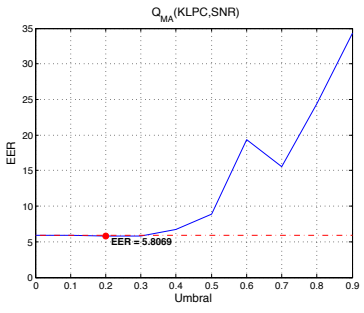

Figura 5.3: Ejemplo de gráfica representando el EER con respecto al umbral de la medida de calidad indicada por el rótulo sobre la figura.

La siguiente figura (figura 5.4) representa las gráficas para cada una de las medidas de calidad combinadas:

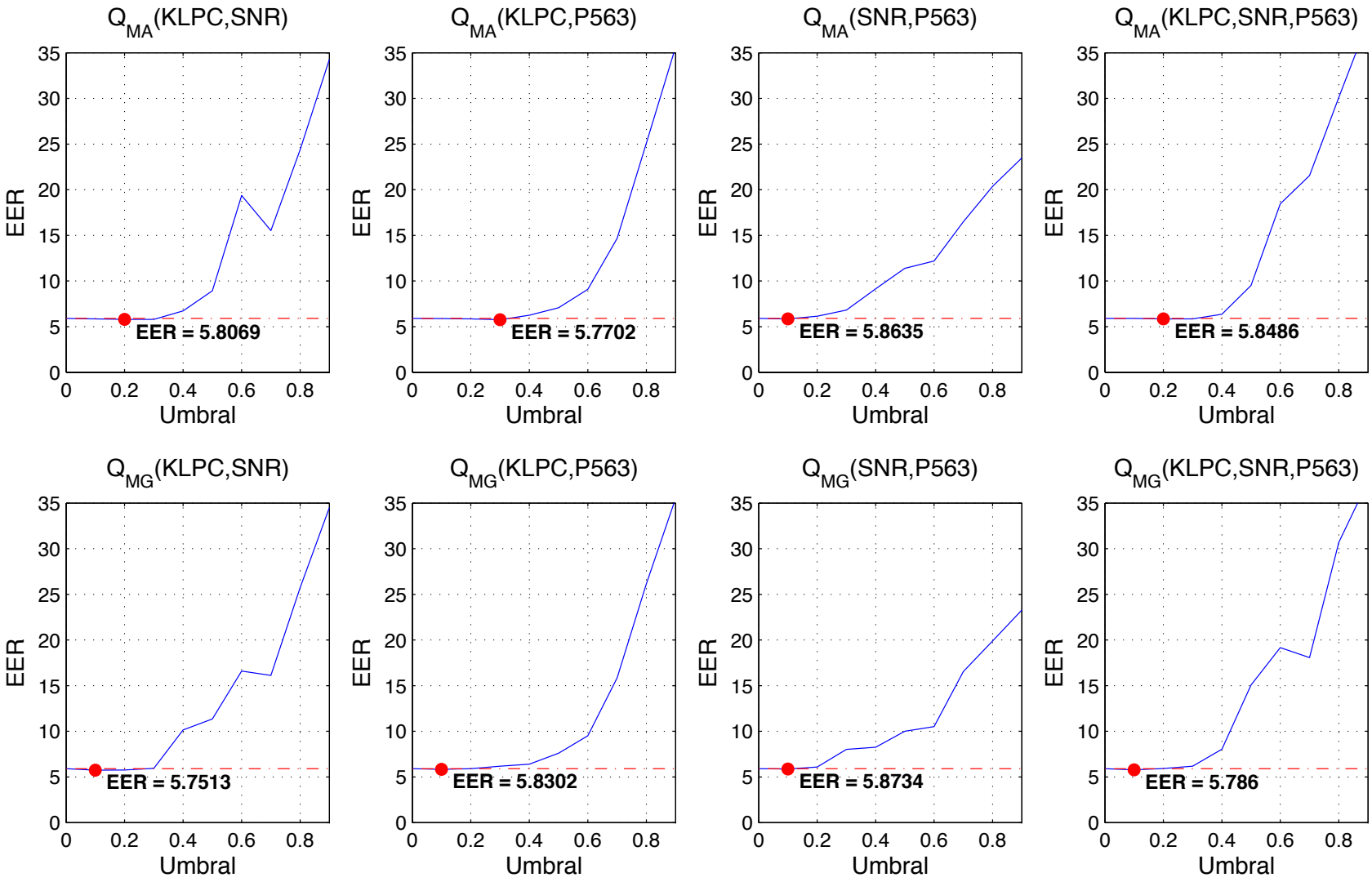

Figura 5.4: Gráficas representando el EER con respecto al umbral de la medida de calidad indicada por el rótulo sobre la figura.

Se puede comprobar que, usando umbrales entre 0.1 y 0.3 obtenemos mejoras de hasta un 2 %. No es una mejora muy significativa, pero únicamente se han usado los ficheros de test. Es de esperar que, una vez se usen los ficheros de train, el EER mejore algo más.

Resultados obtenidos usando locuciones de train y test

En este caso, el umbral de decisión que se ha usado varía entre 0.1 y 0.3 pues son los que mejores resultados han dado en las pruebas anteriores. En las siguientes tablas (tablas 5.2 y 5.3) aparece el EER obtenido en comparación con el que se consiguió usando sólo las etiquetas de los ficheros de test. En verde aparece el mejor EER para cada una de las medidas:

Media Aritmética			
Medida	Umbral	Test	Train-Test
KLPC y SNR	0.1	5.8610	**5.5954**
	0.2	5.8069	5.7597
	0.3	5.8069	5.9474
KLPC y P.563	0.1	5.8953	**5.6187**
	0.2	5.8600	5.8536
	0.3	5.7702	5.9415
SNR y P.563	0.1	5.8635	**5.6917**
	0.2	6.1382	6.2559
	0.3	6.8134	6.9205
KLPC, SNR y P.563	0.1	5.9171	**5.6495**
	0.2	5.8486	5.7126
	0.3	5.8551	5.7612

Cuadro 5.4: Tabla con los diferentes valores de EER usando la la media aritmética sobre locuciones telefónicas.

Media Geométrica			
Medida	Umbral	Test	Train-Test
KLPC y SNR	0.1	**5.7513**	5.8536
	0.2	5.7622	5.8799
	0.3	5.9504	5.8566
KLPC y P.563	0.1	5.8302	**5.8158**
	0.2	5.9018	5.8317
	0.3	6.1853	6.2345
SNR y P.563	0.1	**5.8734**	6.1148
	0.2	6.0911	6.2088
	0.3	8.0304	8.3554
KLPC, SNR y P.563	0.1	5.7860	5.8476
	0.2	5.9266	**5.7831**
	0.3	6.1853	6.3969

Cuadro 5.5: Tabla con los diferentes valores de EER usando la la media geométrica sobre locuciones telefónicas.

Se puede comprobar que las mejoras más significativas se producen con un umbral de 0.1 y más específicamente en las medidas usando la **media aritmética**. En los casos en los que se ha usado la **media geométrica**, el EER no mejora con respecto al uso de medidas de calidad

únicamente en los ficheros de test pero, sin embargo, sigue mejorando con respecto al sistema original en prácticamente todos los casos (salvo en la combinación de SNR y P.563).

Es común que, tras la obtención de los scores en un sistema de verificación, se utilice una normalización de los scores de tipo SNorm para conseguir reducir la variabilidad entre las pruebas con el fin de mejorar aún más el rendimiento de un sistema [22].

El EER obtenido sin utilizar medidas de calidad, pero utilizando la normalización de scores es:

$$EER_{SoloVAD}=\mathbf{5.5720}$$

En este caso, se ha obtenido un valor de EER, aplicando la normalización de los scores, para cada uno de los siguientes casos:

Media Aritmética				
Medida	Umbral	Test	Train-Test	Train-Test-SNorm
KLPC y SNR	0.1	5.6371	**5.3133**	5.3203
	0.2	5.6892	5.4072	5.4072
	0.3	5.7309	5.4792	5.4543
KLPC y P.563	0.1	5.6664	5.4073	5.3838
	0.2	5.5954	**5.3371**	5.3372
	0.3	5.6535	5.572	5.5482
SNR y P.563	0.1	5.7126	5.357	**5.3373**
	0.2	5.8739	5.6773	5.6892
	0.3	6.3218	6.2295	6.2528
KLPC, SNR y P.563	0.1	5.6202	5.29	**5.2726**
	0.2	5.5482	5.3371	5.3369
	0.3	5.6659	5.3471	5.3605

Cuadro 5.6: Tabla con los diferentes valores de EER usando la la media aritmética sobre locuciones telefónicas y aplicando la normalización de los scores (SNorm).

Media Geométrica				
Medida	Umbral	Test	Train	Train-Test-SNorm
KLPC y SNR	0.1	5.5482	**5.3838**	5.4072
	0.2	5.6574	5.4073	5.4166
	0.3	5.7483	5.5482	5.572
KLPC y P.563	0.1	5,7364	5,6892	**5.6694**
	0.2	5,7831	5,6896	5.7126
	0.3	5,7597	5,7364	5.7126
SNR y P.563	0.1	5.642	5.5954	**5.5685**
	0.2	5.8536	5.7831	5.7831
	0.3	7.0513	7.0414	7.028
KLPC, SNR y P.563	0.1	5.6892	5.7111	5.6569
	0.2	5.7637	5.6659	**5.6296**
	0.3	5.8769	5.9008	5.8767

Cuadro 5.7: Tabla con los diferentes valores de EER usando la la media geométrica sobre locuciones telefónicas y aplicando la normalización de los scores (SNorm).

Se puede comprobar que, en el caso de los archivos telefónicos, el uso de la normalización de scores aumenta el rendimiento del sistema y, en la mayoría de los casos, aplicar las medidas de

calidad a las locuciones usadas en la normalización, aumenta aún más el rendimiento, llegando a mejorar hasta un 5.37 %.

5.3.2. Medidas de calidad en locuciones microfónicas

El cálculo de las medidas de calidad de las locuciones microfónicas se ha realizado, al igual que en el caso anterior, sobre las locuciones de test y de train. El rango de valores del umbral de decisión es el mismo que se ha utilizado en los experimentos telefónicos, es decir, los valores entre 0.1 y 0.3. En estos experimentos **se han normalizado los scores**. En este caso, el valor de EER que se obtiene sin utilizar ningún tipo de medida de calidad es:

$$EER_{SoloVAD}\text{=}\mathbf{9.4161}$$

En las siguientes tablas se muestran los valores de rendimiento obtenidos:

Media Aritmética					
Medida	Umbral	Test	Train	Test-Train	Test-Train-SNorm
KLPC y SNR	0.1	14.9866	9.1286	14.3126	14.3114
	0.2	18.2525	8.8411	16.4287	16.2684
	0.3	24.5158	**8.5442**	20.9718	20.5570
KLPC y P.563	0.1	14.7933	9.2229	14.3126	14.2702
	0.2	20.1282	8.9165	19.2139	19.1197
	0.3	25.4583	**8.8411**	23.0642	23.1538
SNR y P.563	0.1	15.9621	9.1522	15.9480	15.9476
	0.2	27.2586	**9.0296**	26.4763	26.3490
	0.3	32.6641	9.4161	31.7970	31.6038
KLPC, SNR y P.563	0.1	14.0487	9.2370	13.5445	13.5162
	0.2	20.5570	9.1286	19.3082	19.3082
	0.3	27.4754	**9.0721**	25.2227	24.9729

Cuadro 5.8: Tabla con los diferentes valores de EER usando la media aritmética sobre locuciones microfónicas y aplicando la normalización de los scores (SNorm).

Media Geométrica					
Medida	Umbral	Test	Train	Test-Train	Test-Train-SNorm
KLPC y SNR	0.1	15.5615	8.9354	14.6991	14.6048
	0.2	20.4628	8.6809	18.5730	18.1677
	0.3	30.2323	**8.6479**	26.8015	26.5140
KLPC y P.563	0.1	22.2866	**9.0296**	21.0377	21.1226
	0.2	26.9004	9.0305	25.9390	25.8400
	0.3	33.8140	9.0394	30.5481	30.4538
SNR y P.563	0.1	19.8360	9.1003	19.0867	19.1102
	0.2	29.4547	**9.0296**	28.9175	28.7243
	0.3	35.8924	9.0721	33.9036	34.0073
KLPC, SNR y P.563	0.1	21.2310	9.0296	20.3638	20.2696
	0.2	27.8571	8.8411	26.0003	26.0568
	0.3	37.7539	**8.6469**	36.0243	36.0252

Cuadro 5.9: Tabla con los diferentes valores de EER usando la media geométrica sobre locuciones microfónicas y aplicando la normalización de los scores (SNorm).

En el caso de las locuciones microfónicas se puede observar que únicamente se producen mejoras cuando se utilizan las medidas de calidad sobre las locuciones de train, produciéndose un empeoramiento de la EER en los demás casos. Las mejoras aun así son muy modestas, llegando a alcanzar, en el mejor de los casos, valores de hasta un 9.25 %.

6

Conclusiones y trabajo futuro

6.1. Conclusiones

En este proyecto se han estudiado distintos métodos de estimación de la calidad de la señal de voz en sistemas de reconocimiento de locutor. A partir del estudio previo de [3] en el que se estudió el impacto que las diferentes medidas tenían sobre un sistema de reconocimiento de locutor, se seleccionaron tres medidas para su estudio en este proyecto: SNR, P.563 y KLPC. En adición a dicho estudio, se decidió obtener nuevas medidas fruto de las diferentes combinaciones entre las medidas de calidad propuestas.

Para realizar este estudio se ha decidido utilizar una base de datos que permita obtener los resultados más actualizados posibles. En este caso se ha elegido la base de datos de NIST, en concreto la perteneciente a la SRE 2012. Además, se han utilizado los dos tipos de canal más comunes (teléfono y micrófono) y se ha estudiado el impacto de las medidas de calidad sobre cada uno por separado (sin realizar experimentos cruzados).

Con respecto al sistema de reconocimiento de locutor, se ha evaluado su rendimiento cuando se hace uso de las medidas de calidad en los distintos tipos de locuciones que intervienen en el sistema. En este caso, las locuciones han sido:

- Locuciones de test.
- Locuciones de entrenamiento (train).
- Locuciones utilizadas para la normalización de scores (SNorm).

Los resultados que se han apreciado sobre las locuciones de tipo telefónicos son bastante satisfactorios, pues se ha demostrado que el uso de las diferentes medidas de calidad sobre los tres tipos de locuciones citadas anteriormente, produciéndose mejoras del rendimiento del sistema **de hasta un 5.27 % relativo**.

En el caso de las locuciones de tipo microfónico, las mejoras no han sido tan evidentes, pues sólo se ha mejorado el rendimiento en los casos en los que se han utilizado las etiquetas de calidad sobre las locuciones de entrenamiento. Algunas de las posibles razones para que sólo haya mejorado con las locuciones de entrenamiento son:

- **Sensibilidad frente a un menor número de muestras de audio**. Es posible que este último sistema sea más sensible a una disminución de las muestras de audio, fruto de la aplicación de las medidas de calidad a las etiquetas del detector de actividad y que presente por ello peores resultados. Como se observó en el capítulo 4, muchas de las locuciones microfónicas presentaban un número muy bajo de muestras una vez se aplicaban las medidas de calidad sobre ellas.

- Otro posible problema es que el cálculo de las medidas de calidad que se ha realizado no sea el más adecuado para este tipo de locuciones, especialmente el de la SNR y la P.563. Observando los datos obtenidos tras la estimación de las medidas de calidad por separado, se puede apreciar que, las locuciones que aparentemente son más ruidosas, presentan un gran número de muestras con un valor de Q muy bajo en los casos de la P.563 y la SNR. Que estas dos medidas presenten el mismo problema se debe a que la correlación entre ellas es muy alta, pues la P.563 utiliza la SNR (entre otros parámetros) para realizar las estimaciones.

Aun así, en el caso mejor, en el que sólo se usan las medidas de calidad sobre las locuciones de entrenamiento, se ha llegado a producir una mejora relativa de **hasta un 9.25 %**.

6.2. Trabajo futuro

Las futuras líneas de trabajo podrían hacer uso de este estudio para mejorar el cálculo de medidas de calidad como la SNR y la P.563 en las locuciones de tipo microfónico paran conseguir mejoras más significativas.

Otra línea de trabajo que podría ser importante consistiría en utilizar otro de los indicadores de degradación que se proponen en [3] y comprobar si, en este caso, también se producen mejoras.

Como se ha mostrado en este proyecto, las locuciones microfónicas presentan mayores dificultades que las puramente telefónicas. Sería necesario realizar un estudio muy exhaustivo de este tipo de locuciones para comprobar qué medidas de calidad son más adecuadas o qué método de cálculo da mejores resultados.

Además, sería importante desarrollar un método de combinación de las medidas de calidad más elaborado que la simple media aritmética o geométrica que permita aumentar el rendimiento de una forma más significativa.

Glosario de acrónimos

- **DET**: Detection Error Tradeoff
- **EER**: Equal Error Rate
- **ITU**: International Telecommunication Union
- **KLPC**: Kurtosis LPC
- **LPC**: Linear Predictive Coding
- **MFCC**: Mel-Frequency Cepstral Coefficients
- **NIST**: National Institute of Standards and Technology
- **SNorm**: Scores Normalization
- **SNR**: Signal-to-Noise Ratio
- **VAD**: Voice Activity Detection

Bibliografía

[1] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems For Video Technology*, 2004.

[2] J. González-Rodríguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-García. Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, n. 7.*, pages 2104–2115, September 2007.

[3] A. Harriero, D. Ramos, J. González-Rodríguez, and J. Fiérrez. Analysis of the utility of classical and novel speech quality measures for speaker verification. *Advances in Biometrics.Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings*, pages 434–442, June 2009.

[4] C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[5] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition Second Edition*. Springer, 2009.

[6] A. K. Jain, P. J. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2008.

[7] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, pages 430–451, 2004.

[8] D. A. Reynolds. An overview of automatic speaker recognition technology. *Proc. International Conference on Acoustics, Speech, and Signal Processing in Orlando*, pages 4072–4075, May 2003.

[9] Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark A. Przybocki. The det curve in assessment of detection task performance. In *EUROSPEECH'97*, 1997.

[10] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[11] D. Benini. Biometric sample quality standards: Importance, status, and direction. In *Proc. NIST Biometric Quality Workshop*, November 7-8 2007.

[12] F. Alonso Fernández. *Biometric Sample Quality and its Application to Multimodal Authentication Systems*. PhD thesis, Universidad Politécnica de Madrid, September 2008.

[13] A. Hicklin and R. Khanna. The role of data quality in biometric systems. *Mitretek Systems*, February 2006.

[14] J. Fiérrez-Aguilar, J. Ortega-García, J. González-Rodríguez, and J. Bigun. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38:777–779, November 2004.

[15] V. Grancharov and W.B. Kleijn. *Speech Quality Assessments*, chapter 5, pages 83–102. Springer, 2007.

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

[17] L. Malfait, J. Berger, and M. Kastner. P.563 - The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, November 2006.

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

[19] Mean Opinion Score (MOS) terminology, ITU-T Rec. P.800.1, 2003.

[20] D. García-Romero, J. Fiérrez-Aguilar, J. González-Rodríguez, and J. Ortega-García. Using quality measures for multilevel speaker recognition, 2005.

[21] NIST SRE 2012 - http://http://www.nist.gov/itl/iad/mig/sre12.cfm.

[22] Zahi N. Karam, William M. Campbell, and Najim Dehak. Towards reduced false-alarms using cohorts. *IEEE ICASSP*, pages 4512–4515, 2011.

A

Presupuesto

1)	**Ejecución Material**	
	▪ Compra de ordenador personal (Software incluido)	2000 €
2)	**Gastos generales**	
	▪ sobre Ejecución Material	352 €
3)	**Beneficio Industrial**	
	▪ sobre Ejecución Material	132 €
4)	**Honorarios Proyecto**	
	▪ 1800 horas a 15 €/ hora	27000 €
5)	**Material fungible**	
	▪ Gastos de impresión	130 €
	▪ Encuadernación	200 €
6)	**Subtotal del presupuesto**	
	▪ Subtotal Presupuesto	29814 €
7)	**I.V.A. aplicable**	
	▪ 21 % Subtotal Presupuesto	6260,94 €
8)	**Total presupuesto**	
	▪ Total Presupuesto	36074,94 €

Madrid, Junio 2014

El Ingeniero Jefe de Proyecto

Fdo.: Pedro Cerame Lardies

Ingeniero Superior de Telecomunicación

B

Pliego de condiciones

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *DETECCIÓN AUTOMÁTICA DE VOZ DEGRADADA USANDO MEDIDAS DE CALIDAD*. 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 Contratać 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.