(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 600 461**

(51) Int. CI.:

G10L 17/06 (2013.01)
G10L 17/20 (2013.01)
G10L 17/04 (2013.01)
G10L 17/10 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **26.04.2013 E 13165466 (7)**

(97) Fecha y número de publicación de la concesión europea: **12.10.2016 EP 2797078**

(54) Título: **Estimación de la fiabilidad del reconocimiento de un orador**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
09.02.2017

(73) Titular/es:

AGNITIO S.L. (100.0%)
Calle Gran Via 39- 8a planta
28013 Madrid, ES

(72) Inventor/es:

BUERA RODRÍGUEZ, LUIS;
VAQUERO AVILÉS-CASCO, CARLOS y
VILLALBA LÓPEZ, JESÚS ANTONIO

(74) Agente/Representante:

MILTENYI, Peter

DESCRIPCIÓN

Estimación de la fiabilidad del reconocimiento de un orador.

Los sistemas de reconocimiento de un orador tienen dos aplicaciones diferentes. Pueden utilizarse para la verificación de un orador, en la que se confirma o se rechaza que una persona que está hablando sea la persona indicada. En este caso, se comparan dos impresiones de voz. La otra aplicación es la identificación de un orador que puede utilizarse para decidir cuál de una serie de personas cuyas impresiones de voz son conocidas para el sistema corresponde a la persona que ha estado hablando. En tales sistemas utilizados para la identificación de un orador, es posible que el orador que está hablando no esté incluido en el conjunto de personas conocidas (conjunto abierto) o puede ser operado de manera que los oradores estén siempre en el conjunto de personas conocidas para el sistema (conjunto cerrado). Normalmente, este tipo de sistemas de reconocimiento de un orador comprenden, para cada orador inscrito en el sistema, un modelo de orador que describe la impresión de voz del orador (la huella de voz que comprende características típicas para el orador).

En sistemas de reconocimiento de un orador actuales, puede ser un problema identificar si el sistema de reconocimiento proporciona decisiones fiables. En particular, en entornos ruidosos o en el caso de falta de correspondencia de canal (siendo un canal todo lo que hay entre la persona que habla y el medio de grabación), los sistemas de reconocimiento de un orador actuales pueden proporcionar resultados poco fiables. Esta falta de correspondencia del canal puede producirse, por ejemplo, si una señal de voz se transmite de una manera que no sea conocida para el sistema y no se ha utilizado para entrenamiento.

Se han realizado varios intentos para superar estos problemas. Ejemplos son las publicaciones de M. C. Huggins y J. J. Grieco: "*Confidence Metrics for Speaker Identification*", publicada en la 7ª ICSLP, Denver, Colorado, 2002, o el documento "*Using Quality Measures for Multilevel Speaker Recognition*", *Computer Speech and Language* 2006; 20 (2 - 3): 192-209 por D. García-Romero, y otros. Se han realizado otros intentos por W. M. Campbell y otros en el documento "*Estimating and Evaluating Confidence for Forensic Speaker Recognition*" en ICASSP 2005; 717-720, y en "*Considering Speech Quality in Speaker Verification Fusion*" en Interspeech 2005 por Y. Solewicz y M. Koppel y los dos documentos de J. Richiardi y otros, titulados "*A Probabilistic Measure of Modality Reliability in Speaker Verification*" en ICASSP, 2005 y el documento "*Confidence and Reliability Measures in Speaker Verification*", publicado en *Journal of the Franklin Institute* 2006; 343 (6): 574-595.

En algunos de estos enfoques se utilizan Redes Bayesianas (RB). Un documento que puede ayudar a entender las Redes Bayesianas es, por ejemplo, "*Pattern Recognition and Machine Learning*" de C. Bishop, publicado en *Springer Science and Business Media*, LLC, 2006.

Una Red Bayesiana es un modelo gráfico probabilístico que representa un conjunto de variables (aleatorias) y sus dependencias condicionales. Sus nodos pueden representar una o más de las variables vistas y/u ocultas y/o hipótesis y/o parámetros determinísticos.

Una variable en función de otra variable será representada en una Red Bayesiana por una flecha que apunta desde la primera variable (variable padre), de que la depende la segunda variable (variable hija), hacia la segunda variable (dependiente).

Una red de este tipo puede ser entrenada. Con una red de este tipo (entrenada), dado un conjunto de parámetros conocidos (vistos), puede estimarse la probabilidad de una variable oculta.

Trabajos previos sobre la fiabilidad basados en Redes Bayesianas pueden tener el inconveniente de que los parámetros de la Red Bayesiana pueden depender del umbral (punto de trabajo) de reconocimiento de un orador tal como, por ejemplo, en la publicación de Richiardi y otros en ICASSP '05. En ese caso, una modificación del punto de trabajo requeriría un entreno nuevo y completo de la Red Bayesiana.

Otros problemas que pueden estar presentes en la técnica anterior son, por ejemplo, el hecho de que la degradación de la señal puede afectar a la fiabilidad de la prueba de manera diferente si la prueba es objetivo o no objetivo y/o que para el proceso de entreno pueden ser necesarias realizaciones limpias y degradadas de las mismas expresiones (lo cual se denomina también datos estéreo). En particular, esto puede significar que, para entrenar a los sistemas de la técnica anterior, puede ser necesario contar con las expresiones de entreno como señales con y sin distorsiones, por ejemplo, causadas por canales, estrés del orador, calidad de datos, convolución, ruido añadido u otras influencias que degradan los datos. Todos estos datos no son siempre fáciles de proporcionar y, a veces, se desconoce la correlación entre la fiabilidad y la distorsión de la señal.

Finalmente, la técnica anterior ha demostrado que la fiabilidad de una prueba (comparación entre un audio de prueba y un modelo de orador) es más profunda relacionada con la calidad de señal tanto del (de los) de audio(s) de

prueba como del (de los) audio(s) modelo que con una calidad de señal individual de audio(s) de prueba o audio(s) modelo solamente. Un modelo de orador, tal como se utiliza en este texto, generalmente está constituido por un sistema de reconocimiento de un orador que utiliza uno, dos, tres o más audios modelo.

Otra técnica anterior relevante se encuentra en J. Villalba y otros "*Reliability Estimation of the Speaker Verification Decisions Usign Bayesian Networtks to Combine Information from Multiple Speech Quality Measures*", *Advances in Speech and Language Technologies for Iberian Languages*, vol. 328, páginas 1-10, 21 de noviembre de 2012.

Los objetivos de la presente invención son determinar y proporcionar un valor que describa la fiabilidad de una prueba determinada, modelar el efecto de medidas de calidad en el rango las calificaciones, no sólo en un punto de trabajo (en el que un punto de trabajo puede ser, por ejemplo, el umbral de reconocimiento de un orador del (de los) sistema(s) de reconocimiento de un orador, que puede afectar al proceso de entreno de la solución), y o el uso de (tanta) información útil (como sea posible) a partir de la señal de voz o cualquier otra fuente y/o evitar el uso de datos estéreo durante el entreno de la Red Bayesiana (que comprende la señal sin degradación y la señal con degradación, por ejemplo, efectos del canal, pérdidas de calidad de datos, etc.).

La invención pretende superar por lo menos algunos de los problemas mencionados anteriormente. En particular, los resultados para la fiabilidad de una prueba determinada pueden ser mejores si la calidad del audio de la prueba y el modelo de orador no se consideran de manera independiente.

La invención comprende un procedimiento de acuerdo con la reivindicación 1, un medio legible por ordenador de acuerdo con la reivindicación 12 y un sistema de acuerdo con la reivindicación 13.

Un sistema de reconocimiento de un orador puede utilizar un modelo de orador basado en uno, dos, tres o más audio(s) modelo para calcular un resultado relativo a un audio de prueba. Este resultado puede utilizarse en un procedimiento de acuerdo con la invención.

Un audio de prueba suele ser una señal de voz (de prueba) que comprende, por ejemplo, una expresión o un fragmento de una expresión, por ejemplo, una palabra o frase de un orador. El (los) audio(s) modelo generalmente es (son) uno, dos, tres o más señales de voz (inscripción) que comprenden cada una, por ejemplo, una expresión o un fragmento de una expresión, por ejemplo, una palabra o frase de un orador.

Generalmente, se utiliza un modelo de orador basado en uno, dos, tres, o más audios modelo en combinación con un audio de prueba en un sistema de reconocimiento de un orador. Sin embargo, también es posible utilizar un modelo de orador basado en uno, dos, tres o más audios modelo y compararlo con una impresión de voz de prueba en base a uno, dos, tres, o más audios de prueba (o directamente con uno, dos, tres, o más audios de prueba).

Si se utilizan dos, tres, o más audios prueba, cada audio de prueba puede ser analizado por el sistema de reconocimiento de un orador independientemente y los resultados puede fusionarse entonces, por ejemplo, promediando el resultado. En otros casos, puede extraerse primero una impresión de voz de prueba de audios de prueba (ya que normalmente se crea un modelo de orador a partir del uno, dos, tres, o más audios modelo). Este enfoque general proporciona resultados más fiables.

Siempre que en el texto siguiente se mencione "una prueba de audio" o "la prueba de audio", esto puede referirse a una prueba de audio o dos, tres, o más de pruebas de audio o una impresión de voz de prueba generada a partir de una, dos, tres, o más pruebas de audio.

La invención comprende un procedimiento para estimar la fiabilidad de un resultado de un sistema de reconocimiento de un orador, refiriéndose el resultado a un audio de prueba y un modelo de orador, que se basa en uno, dos, tres o más audio(s) modelo, utilizando el procedimiento una Red Bayesiana para estimar si el resultado es fiable.

El sistema de reconocimiento de un orador puede haber utilizado el uno, dos, tres o más audio(s) modelo para construir un modelo de orador basado en el uno, dos, tres o más de audio(s) modelo antes de calcular el resultado relativo al audio de prueba y el uno, dos, tres o más de audio(s) modelo. El modelo de orador generalmente describe la impresión de voz del orador una vez que ha sido entrenado o creado en base al uno, dos, tres o más audio(s) modelo. En algunas realizaciones, el modelo de orador puede ser un audio modelo o puede corresponder a uno, dos, tres o más de audio(s) modelo.

La fiabilidad puede correlacionarse con algunas medidas de calidad que pueden derivarse directa o indirectamente de las señales de voz (por ejemplo, audio(s) modelo y/o audio(s) prueba). Estas medidas de calidad pueden elegirse para comprender información sobre el audio de prueba y/o el modelo de orador (audio(s) modelo). Normalmente, las

medidas de calidad deben describir tanta información como sea posible sobre el audio de prueba y/o el orador modelo.

Para estimar la fiabilidad de la decisión del sistema de reconocimiento de un orador se utiliza una, dos, tres, cuatro, o más de cuatro medidas de la calidad del audio de prueba y una, dos, tres, cuatro, o más de cuatro medidas de la calidad del (de los) audio(s) modelo. En general, para cada uno del (de los) audio(s) modelo se determina por separado una, dos, tres, cuatro, o más de cuatro medidas de calidad. En otras realizaciones, la una, dos, tres, cuatro, o más de cuatro medidas de calidad puede determinarse a partir del modelo de orador. Las una, dos, tres, cuatro, o más medidas de calidad del audio de prueba suelen ser las mismas medidas de calidad que la una, dos, tres, cuatro o más medidas de calidad del (de los) audio(s) modelo. Si P es el número de medidas de calidad (una, dos, tres, cuatro o más) después, de cada modelo y prueba de audio puede extraerse medidas de calidad P.

Después, todas las medidas de calidad extraídas de los audios modelo y el audio de prueba pueden ser incluidas directamente en la Red Bayesiana. En ese caso, el número de variables de calidad sería P veces el número de audios implicados (la suma de los audios modelo utilizados aquí y el número de audios de prueba implicados). Tal como se ha indicado anteriormente, puede haber uno, dos, tres o más audios modelo y uno, dos, tres o más audios de prueba.

Alternativamente, las medidas de calidad de todos los audios modelo pueden comprimirse como si se originaran de un audio modelo y pueden incluirse entonces en la Red Bayesiana de esa manera. A partir del audio de prueba, puede derivarse el mismo número de medidas de calidad.

Si se utilizan dos, tres, cuatro o más pruebas de audio, sus medidas de calidad pueden comprimirse de la misma manera que las medidas de calidad de los audios modelo antes de que las variables que describen las medidas de calidad P sean utilizadas por la Red Bayesiana. En ese caso, se utilizarían dos veces el número de medidas de calidad utilizadas: variables (2P) que describen medidas de calidad (P describiendo los audios modelo y P describiendo las pruebas de audio) como variables para la Red Bayesiana.

Alternativamente, puede utilizarse una combinación de estos dos procedimientos descritos anteriormente de introducir las medidas de calidad en la Red Bayesiana: por ejemplo, sería posible que, aunque las medidas de calidad se deriven de cada audio modelo y se utilicen directamente en la Red Bayesiana, si se utilizan dos, tres, cuatro, o más audios de prueba, sus medidas de calidad pueden comprimirse antes de utilizarse, de modo que el número de medidas de calidad utilizadas en la Red Bayesiana puede ser P veces el número de audios modelo más uno. Alternativamente, las medidas de calidad de los audios modelo pueden comprimirse mientras que las medidas de calidad de dos, tres, o más audios de prueba pueden utilizarse sin comprimir.

En un caso alternativo, las medidas de calidad del (de los) audio(s) de prueba y el (los) audio(s) modelo pueden comprimirse entre sí de manera que sólo se utiliza una vez el número de variables P de medidas de calidad como entrada para la Red Bayesiana (como si solamente estuviera presente el conjunto de medidas de calidad).

Las medidas de calidad pueden derivarse del audio de prueba y el (los) audio(s) modelo directamente y/o indirectamente.

La expresión "estimar" puede utilizarse en este contexto debido a que el cálculo de la fiabilidad puede depender, por ejemplo, del entreno de la Red Bayesiana. Sin embargo, dos Redes Bayesianas diferentes, que han sido entrenadas de la misma manera y con los mismos parámetros con las mismas dependencias, generalmente proporcionarán (calcularán) los mismos resultados para la fiabilidad de la misma prueba y el audio modelo (prueba). Por lo tanto, el resultado de la estimación generalmente es no arbitrario.

El resultado del sistema de reconocimiento de un orador puede ser una calificación vista $\hat{s}_i$ proporcionada por el sistema de reconocimiento de un orador. Este resultado visto, por ejemplo, puede ser una calificación o una decisión normalizada, o puede ser un coeficiente de probabilidad o un coeficiente de probabilidad de registro o una calificación en un cierto rango, cada uno de los cuales puede estar opcionalmente calibrado. Un índice *i* puede indicar la prueba *i*-ésima en todo el texto. Aquí, una prueba puede corresponder a comparar un modelo de orador y un audio de prueba. El sistema de reconocimiento de un orador generalmente proporciona una calificación por prueba. Entonces, la calificación resultante del sistema de reconocimiento de un orador y las medidas de calidad derivadas del modelo de orador (el uno, dos, tres, cuatro, o más audios modelo) y el audio de prueba se utilizan como parámetros de entrada para la Red Bayesiana y son procesados por la Red Bayesiana.

En otros casos, el resultado del sistema de reconocimiento de un orador puede obtenerse utilizando la calificación vista $\hat{s}_i$ proporcionada por el sistema de reconocimiento de un orador, y comparando la calificación vista $\hat{s}_i$ contra un umbral de reconocimiento de un orador. De esta manera, puede proporcionarse un resultado booleano. El resultado siempre suele ser "verdadero" si la calificación es mayor que el umbral de reconocimiento de un orador y "falso" si la

calificación es inferior al umbral de reconocimiento de un orador. Para resultados iguales al umbral de reconocimiento de un orador, el resultado puede definirse como "verdadero" o "falso" en un sistema de reconocimiento de un orador.

En particular, un procedimiento para estimar la fiabilidad de una decisión de un sistema de reconocimiento de un orador puede utilizar una Red Bayesiana para estimar la distribución posterior de una calificación oculta s_i dada la calificación vista $\hat{s}_i$ y las medidas de calidad. En este documento, $\hat{s}$ (la calificación vista) y s (la calificación oculta) son variables aleatorias. Cada $\hat{s}_i$ y s_i es una realización de estas variables aleatorias. Aquí, cada $\hat{s}_i$ es una realización de $\hat{s}$, y de manera correspondiente cada s_i es una realización de s.

La calificación oculta s_i es la calificación que se obtendría si ninguna degradación hubiera afectado el audio de pruebas y el (los) audio(s) modelo (señales de voz de inscripción). La calificación oculta s_i también se conoce como calificación limpia s_i en este texto.

La calificación vista $\hat{s}_i$ es la calificación proporcionada por el sistema de reconocimiento de un orador con la prueba y el (los) audio(s) modelo reales. El (los) audio(s) de prueba y el (los) audio(s) modelo también pueden referirse a señales de voz en este texto cuando está claro de qué señal(es) de voz se trata. Partes o todas las señales de voz (audio(s) de prueba y/o de audio(s) modelo) para los que el sistema de reconocimiento de un orador proporciona el resultado pueden haber sido degradadas.

Tal estimación de la fiabilidad puede permitir, por ejemplo, eliminar la dependencia del punto de trabajo de la Red Bayesiana que se produce en otros enfoques de la técnica anterior cuando se obtiene la fiabilidad para una prueba.

En un procedimiento tal como se ha descrito anteriormente, puede suponerse que las calificaciones ocultas y vistas están relacionadas linealmente, por ejemplo, que, $\hat{s}_i = s_i + \Delta s_i$ en el que Δs_i describe la desviación (diferencia) entre la calificación vista y oculta.

La desviación (diferencia) Δs_i entre la calificación vista y oculta sigue una distribución definida por uno, dos o más parámetros. Estos parámetros generalmente dependen del estado de la calidad y la naturaleza de la prueba. Esta distribución de la diferencia suele ser diferente en función de las medidas de calidad. Por ejemplo, para diferentes relaciones de señal - ruido, la distribución de la diferencia no suele ser igual. Además, esto normalmente se ve influenciado en función de si la prueba es objetivo o no objetivo.

La Red Bayesiana puede utilizarse como nodos que describen parámetros vistos de las N calificaciones vistas $\hat{s}_i$ y/o los P grupos de medidas de calidad Q_{pi} (en el que P es el número de medidas de calidad y puede ser 1, 2, 3, 4 o más, $p \in [1, P]$).

El índice i indica la prueba i-ésima. Puede haber N pruebas, de manera que i puede ser un número entre 1 y N ($i \in [1, N]$)- N puede ser 1, 2, 3 o más.

Las medidas de calidad Q_{pi} pueden ser independientes unas de otras dados ciertos estados de calidad z_i. Los estados de calidad z son una variable aleatoria. Cada z_i Es una realización de z. De este modo, puede forzarse la independencia entre variables que deben ser independientes entre sí.

La Red Bayesiana puede utilizar como nodos que describen parámetros ocultos uno, dos, tres, o más, o todos de los siguientes:

- calificación(es) oculta(s) s_i,
- un carácter objetivo o no objetivo de la calificación para cada calificación oculta, θ_i, (también denominado etiqueta de prueba real (oculta) o etiqueta real (oculta) (de la prueba),
- estados de calidad (estados de la calidad) z_i
- coeficientes π_z de la distribución discreta que describen los estados de calidad, z;
- media $\mu_{\Delta s}$ y precisión, $\Lambda_{\Delta s}$, que describen la distribución (opcionalmente Gaussiana) de la distribución de la desviación entre las calificaciones vistas y ocultas, en el que la media y la precisión de Δs describen la variable aleatoria Δs; cada Δs_i es una realización de Δs
- una media μ_s y precisión, Λ_s, que describen la distribución (opcionalmente Gaussiana) de la distribución de la(s) calificación(es) oculta(s) del sistema de reconocimiento de un orador, en el que la media y la precisión describen, de nuevo, la variable aleatoria s; cada s_i sería una realización de esa variable aleatoria
- los coeficientes π_z que describen la distribución de la variable aleatoria z que se aplica para cada prueba i (y, por lo tanto, es independiente de i); π_z puede verse como un vector dimensional K, en el que K es el número de estados de calidad. π_z puede tener elementos π_{zk} ($k \in [1, K]$) que describen la distribución opcionalmente discreta que describe z;

- las distribuciones de los grupos de medidas de calidad (opcionalmente Gaussianas), en el que normalmente hay una media μ_{Qp} y precisión Λ_{Qp} para cada uno de los P grupos de calidad, Q_p en el que p puede ser entre 1 y P $\in$ [1, P]; tales distribuciones de calidad pueden depender del estado de calidad, z, de modo que habrá K diferentes distribuciones; cada Q_{pi} es una realización de Q_p;

Los estados de calidad pueden ser, por ejemplo, un vector binario K-dimensional con elementos z_{ik}, en el que K es el número de estados de calidad y k puede ser entre 1 y K (k $\in$ [1, k]) y K puede ser 1, 2, 3 o más; z_{ik} generalmente es una variable booleana que puede ser 0 ó 1. Los estados de calidad generalmente son estados definidos por ciertos rangos de valores para las medidas de calidad.

Por ejemplo, dadas dos medidas de calidad, que tienen cada una valores entre - ∞ y + ∞ (por ejemplo, relación señal-ruido entre -∞ dB y + ∞ dB y probabilidad de registro UBM entre -∞ y ∞), entonces es posible definir varios estados de calidad.

Por ejemplo, dadas las medidas de calidad de este ejemplo, puede definirse un estado de calidad mediante la primera variable que sea menor de 15 dB y una segunda variable que sea menor de 0. Esto puede escribirse como que z es el vector (1, 0, 0, 0). El segundo estado de calidad puede definirse por la primera variable que sea inferior de 15 dB y la segunda variable que sea mayor o igual que 0 (z = (0, 1, 0, 0)). Como tercer estado de calidad z = (0, 0, 1, 0), la primera variable puede ser mayor o igual que 15 dB y la segunda menor que 0. El cuarto estado de calidad puede venir dado por la primera variable que sea mayor o igual que 15 dB y la segunda variable que sea mayor o igual que 0 (z = (0, 0, 0, 1)). En ese caso, por ejemplo, el número de estados de calidad K sería 4. El número de estados de calidad K se define dependiendo de qué condiciones se elijan para definir estos estados de calidad. La Red Bayesiana también puede utilizar un valor determinístico, es decir, la hipótesis previa, como un nodo. La hipótesis puede escribirse como π_θ, por ejemplo, como $\pi_\theta = (P_T, P_{NT})$ con $P_T + P_{NT} = 1$ y puede ser determinístico. Aquí, P_T puede ser el objetivo previo y P_{NT} el no objetivo previo.

$\hat{s}_i$ puede ser una variable vista dependiente de z_i, s_i, θ_i, $\mu_{\Delta s}$ y/o $\Lambda_{\Delta s}$. θ_i, puede ser dependiente de la π_θ (opcionalmente determinístico), s_i puede depender de θ_i, Λ_s y/o μ_s mientras que μ_s puede depender de Λ_s. z_i puede depender de π_z, μ_{Qp} y/o Λ_{Qp}, mientras que μ_{Qp} puede depender de Λ_{Qp}. Opcionalmente, puede no haber otras dependencias entre los parámetros aparte de las dependencias mencionadas anteriormente.

A partir de las distribuciones estimadas posteriores que se encuentran mediante la Red Bayesiana, la probabilidad de la fiabilidad puede calcularse teniendo en cuenta la calificación vista $\hat{s}_i$ y las medidas de calidad correspondientes a la calificación vista (lo que normalmente significa las medidas de calidad asociadas al audio de prueba y al (a los) audio(s) modelo a partir de las cuales se ha calculado la calificación vista $\hat{s}_i$ mediante el sistema de reconocimiento de un orador).

Normalmente, las medidas de calidad también se vistas. Las medidas de calidad pueden derivarse directamente de las señales de voz (de prueba y/o de inscripción) (audio de prueba y/o audio(s) modelo). En otras realizaciones, la medida de calidad puede derivarse de las señales de voz parcialmente o completamente de manera indirecta y/o parcialmente de manera directa.

Puede tomarse entonces una decisión definitiva de si la prueba es fiable o no utilizando un umbral de fiabilidad.

La fiabilidad en este contexto puede definirse tal como sigue.

Para una prueba, i, un sistema de reconocimiento de un orador puede tomar la decisión $\hat{\theta}_i$, a saber, puede decidir que la persona es el objetivo (lo que significa que el audio de prueba lo dijo la persona supuesta) si la calificación vista $\hat{s}_i$ es mayor o igual que el umbral de reconocimiento de un orador del sistema de reconocimiento de un orador ($\hat{s}_i \geq \varphi_\theta$). Si la calificación vista $\hat{s}_i$ es menor que el umbral de reconocimiento de un orador del sistema de reconocimiento de un orador φ_θ ($\hat{s}_i < \varphi_\theta$), se supone que la persona no es el objetivo. Esto también puede escribirse como

$$\hat{\theta}_i = \begin{cases} T & if \quad \hat{s}_i \geq \varphi_\theta \\ NT & if \quad \hat{s}_i < \varphi_\theta \end{cases}$$

en el que T es la decisión de que la persona es el objetivo y NT es la decisión de que la persona no es el objetivo. Alternativamente, puede decidirse que una persona es el objetivo si el resultado visto es mayor que el umbral de reconocimiento de un orador y no es el objetivo si la calificación vista es menor o igual que el umbral de reconocimiento de un orador.

Para determinar si la decisión $\hat{\theta}_i$ es fiable, puede calcularse una distribución posterior (P (s_i | $\hat{s}_i$, Q_i)) de la calificación hipotética oculta s_i con la Red Bayesiana propuesta dada la calificación vista $\hat{s}_i$ y las medidas de calidad Q_i. Los valores para la distribución posterior de la calificación hipotética oculta s_i dada $\hat{s}_i$ y las medidas de calidad Q_i pueden calcularse, por ejemplo, tal como se describe en el Anexo I.

Normalmente Q es una variable aleatoria de la medida de la calidad. Se trata, normalmente, de un vector con P elementos, un elemento por medida de la calidad. Q_i es una realización de Q para la prueba i-ésima. Por lo tanto, normalmente Q_i es también un vector de P elementos. Cada elemento de Q_i puede denominarse Q_{pi}.

Entonces, la probabilidad de fiabilidad puede calcularse, por ejemplo, utilizando la siguiente expresión:

$$P(R_i = \Re \mid \hat{s}_i, Q_i) = \begin{cases} P(s_i \geq \varphi_\theta \mid \hat{s}_i, Q_i) = \int_{\varphi_\theta}^{\infty} P(s_i \mid \hat{s}_i, Q_i) ds_i & if \quad \hat{\theta}_i = T \\ P(s_i < \varphi_\theta \mid \hat{s}_i, Q_i) = \int_{-\infty}^{\varphi_\theta} P(s_i \mid \hat{s}_i, Q_i) ds_i & if \quad \hat{\theta}_i = NT \end{cases}$$

Puede decidirse entonces tomar la decisión $\hat{R}_i$ si la prueba es fiable $\Re$ o no fiable U, utilizando un umbral de fiabilidad φ_R. $\hat{R}_i$ se considera fiable si la probabilidad de fiabilidad es mayor o igual que el umbral de fiabilidad φ_R y $\hat{R}_i$ se considera no fiable si la probabilidad de fiabilidad es menor que el umbral de probabilidad φ_R.

Esto también puede escribirse, por ejemplo, como

$$\hat{R}_i = \begin{cases} \Re & if \quad P(R_i = \Re \mid \hat{s}_i, Q_i) \geq \varphi_R \\ U & if \quad P(R_i = \Re \mid \hat{s}_i, Q_i) < \varphi_R \end{cases},$$

donde las variables tienen el significado que se ha descrito anteriormente.

Alternativamente, la decisión puede considerarse fiable $\Re$ si la probabilidad de fiabilidad es mayor que el umbral de fiabilidad φ_R y $\hat{R}_i$ se considera no fiable si la probabilidad de fiabilidad es menor que o igual que el umbral de probabilidad φ_R.

Por ejemplo, una prueba puede considerarse fiable si la calificación vista es considerada objetivo (por ejemplo, la calificación vista del sistema de reconocimiento de un orador es superior al umbral de reconocimiento de un orador) y la probabilidad de la calificación oculta es mayor (y opcionalmente igual) que el umbral de reconocimiento de un orador dadas las medidas de calidad y la calificación vista es mayor que un umbral de fiabilidad dado. Puede considerarse una prueba fiable también si el resultado visto se considera que es no objetivo (por ejemplo, la calificación vista del sistema de reconocimiento de un orador es inferior al umbral de reconocimiento de un orador) y la probabilidad de la calificación oculta es inferior al umbral de reconocimiento de un orador dadas las medidas de calidad y la calificación vista es mayor (y opcionalmente igual) que un umbral de fiabilidad dado.

De lo contrario, puede considerarse no fiable.

La Red Bayesiana utilizada en un procedimiento descrito anteriormente puede entrenarse antes de que ser utilizada para estimar la fiabilidad de un resultado de un sistema de reconocimiento de un orador. Para el entreno de la Red Bayesiana, pueden utilizarse varias señales de voz de inscripción. Pueden corresponder o no parcialmente a las señales de voz de inscripción. En particular, sus parámetros pueden entrenarse, por ejemplo, utilizando varias señales de voz de desarrollo de uno, dos, tres, o más de tres, en particular, más de diez y, en particular, más de 50 oradores diferentes. En particular, para el entreno pueden utilizarse más de 20, por ejemplo, más de 50, por ejemplo, más de 100 señales de voz de desarrollo.

Además, estas señales de voz de desarrollo tienen preferiblemente una alta variabilidad en términos de medidas de calidad de la señal, ya que cualquier relación (oculta) entre fiabilidad y medidas de la calidad de la señal no vistas durante el entreno de la Red Bayesiana generalmente después no se modelan.

En particular, las señales de voz de desarrollo pueden grabarse o recibirse así de dos o más canales diferentes, por ejemplo, teléfonos y ordenadores. También pueden comprender uno, dos, tres o más deformaciones de calidad diferentes (por ejemplo, ruido de fondo, deformación de la señal debido a error del portador de datos, ruido

aleatorio). Preferiblemente, en diferentes señales de voz de desarrollo hay presentes dos, tres, cuatro, o más degradaciones de datos diferentes.

Para el entrenamiento de la Red Bayesiana y/o para la estimación de la fiabilidad de una decisión de un sistema de reconocimiento de un orador pueden utilizarse, una, dos, tres, cuatro, o más de cuatro medidas de calidad. El uso de dos, tres, cuatro, o más medidas de calidad puede ser ventajoso dado que puede permitir tener en cuenta diferentes tipos de degradación de datos permitiendo estimar, por lo tanto, por ejemplo, las fiabilidades de más señales correctamente.

En dicho procedimiento, las medidas de calidad que se utilizan para el entrenamiento de la Red Bayesiana y/o para la estimación de la fiabilidad de la decisión de un sistema de reconocimiento de un orador pueden comprender la relación señal ruido (SNR) y o el índice de modulación (MI) y o la entropía y o la probabilidad de registro de modelo de fondo universal (UBMLLK).

En particular, la relación señal ruido puede utilizar propiedades de intervalos de voz hablada. Mientras que la mayor parte de la energía de la voz hablada puede concentrarse en múltiplos de su frecuencia de tono, ruidos aditivos pueden tener una distribución de frecuencia más uniforme. Esto puede permitir el uso de filtros de peine adaptados dependientes del tiempo para estimar una señal limpia y una potencia de ruido por separado en segmentos de voz para cada trama, donde una trama es un pequeño pedazo de audio extraído mediante ventanas (por ejemplo, podrían utilizarse ventanas *Hamming* o *Hanning* para tal fin). Por ejemplo, unas proporciones de segmentos de voz en el habla pueden ser suficientemente elevadas para seguir la evolución del ruido en una amplia gama de aplicaciones reales y proporcionar una medida marco a marco. Estas mediciones pueden calcularse en un dominio de tiempo corto, de manera que la señal de voz se divide en trozos pequeños denominados tramas, utilizando algún tipo de ventanas tales como por ejemplo *Hamming* o *Hanning*.

Este enfoque puede ser más robusto que otros enfoques que utilizan segmentos de silencio para estimar la potencia de ruido, especialmente contra los ruidos no estacionarios. Sin embargo, en otras realizaciones, los segmentos de silencio pueden utilizarse para estimar la potencia de ruido, o pueden utilizarse otros procedimientos para estimar la relación señal ruido.

J. Villalba y otros describen "*Comb Filter Ratio from Local SNR Estimation v1.2*", en un informe técnico de la Universidad de Zaragoza, Zaragoza, España, 2009.

El índice de modulación, por ejemplo, que se explica en el documento de J. Villalba, "*Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems*", publicado en COST 2011 *European Workshop, BiolD* 2011, Brandenburg, 2011, pp. 274 - 285, Springer Berlin/Heidelberg, puede extraerse de la envolvente de la señal, por ejemplo, en una ventana especificada mediante el cálculo de la relación de la diferencia de los valores máximos y mínimos y la adición de los valores máximos y mínimos de la envolvente. Suponiendo que la voz es una señal modulada, algunas distorsiones tales como ruido aditivo o distorsiones convolucionales pueden modificar esta métrica.

La entropía (que se explica, por ejemplo, en J. Villalba y otros, "*Quality Measure Fusion v1.0 and Entropy Measure v1.0"*, Informe Técnico de la Universidad de Zaragoza, Zaragoza, España, 2009) puede medir la incertidumbre sobre la realización de una variable aleatoria. La entropía relacionada con una variable aleatoria X con L observaciones $\{o_1, o_2 ..., o_L\}$ y las distribuciones de probabilidad $\{p_1, p_2 ..., p_l\}$ pueden definirse como

$$H(X) = \sum_{l=1}^{N} p_l \log(p_l)$$

La probabilidad de registro de modelo de fondo universal (UBMLLK) (tal como se explica, por ejemplo, en el documento "*Reliability Estimation from Quality Measures in Speaker Verification II*"; Informe Técnico, Universidad de Zaragoza, Zaragoza, España, 2011 por J. Villalba y E. Lleida y en el documento "*Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification*", en: M. Tistarelli, M. Nixon eds. *Advances in Biometrics*. Vol. 5558, Springer Berlin/Heidelberg; 2009:434 - 442 por A. Harriero y otros) puede indicar la degradación de una expresión de voz en términos con su divergencia respecto a un modelo de fondo universal determinado (UBM). Tal UBM puede ser el punto de partida del sistema de reconocimiento de un orador. Por lo tanto, una expresión que esté bien representada por el UBM (alto UBMLLK) puede alcanzar una precisión satisfactoria.

También pueden utilizarse una o más de otras medidas de calidad en lugar de una o más, o además de las una o más, de las medidas de calidad mencionadas anteriormente.

La Red Bayesiana puede operar y/o ser entrenada de manera independiente del umbral de reconocimiento de un orador del sistema de reconocimiento de un orador. Por lo tanto, un mero cambio del umbral de reconocimiento de un orador del sistema de reconocimiento de un orador puede no hacer que sea necesario un re-entreno de la Red Bayesiana.

Durante el entreno de la Red Bayesiana, pueden verse una o más de las variables ocultas, por ejemplo, ser conocida por las señales de voz de desarrollo correspondientes.

Para el entreno de la Red Bayesiana, puede utilizarse un algoritmo de Maximización de Expectativa (EM) para extraer los parámetros del modelo completo.

Una descripción del algoritmo de Maximización de Expectativa se da, por ejemplo, en el documento "*Maximum likelihood from incomplete data via EM algorithm*", por A. P. Dempster, en el *Journal of Royal Statistical Society*, 39 (1): 1-38.

El modelo puede ser, por ejemplo, un modelo M = (μ_s, Λ_s, μ_{Qp}, Λ_{Qp}, π_z, $\mu_{\Delta sk\theta}$, $\Lambda_{\Delta sk\theta}$). Los parámetros del modelo pueden ser μ_S y Λ_s siendo métricas de media y de precisión de la distribución (opcionalmente Gaussiana) que describen s_i, μ_{Qp} y Λ_{Qp} siendo métricas de media y de precisión de la distribución (opcionalmente Gaussiana) que describen los grupos vistos de medidas de calidad Q_p, π_z siendo los coeficientes de la distribución opcionalmente discreta que describen z, $\mu_{\Delta sk\theta}$ y $\Lambda_{\Delta sk\theta}$ siendo métricas de media y de precisión de la distribución (opcionalmente Gaussiana) que describen $\Lambda_{sk\theta}$, que es la desviación entre calificaciones vistas y ocultas de un determinado θ (objetivo o no objetivo) y un estado de calidad determinado (k). Estos parámetros pueden corresponder, por ejemplo, a los parámetros de la Red Bayesiana descritos anteriormente.

En una Red Bayesiana, puede haber P grupos de nodos que comprenden Q_p, μ_{Qp} y Λ_{Qp} donde P puede ser el número de medidas de calidad.

También puede haber N grupos de nodos que comprenden $\hat{s}_i$, s_i, Q_{pi}, θ_i y z_i. Aquí, N puede ser el número de pruebas.

La Red Bayesiana puede ser entrenada de una manera supervisada, de una manera no supervisada o de una manera ciega.

En una manera supervisada, Δs y z son vistas en el entreno y ocultas en las pruebas.

Entonces $P(\Delta s \mid z_k = 1, \theta)$ y $P(Q \mid z_k = 1)$

puede construirse para cada posible combinación de las medidas de calidad. Aquí, $P(\Delta s \mid z_k = 1, \theta)$ es la distribución de probabilidad de la desviación entre calificaciones vistas y ocultas dado un cierto estado de calidad (k) y un cierto tipo de prueba, a saber, objetivo o no objetivo (θ). La segunda expresión $P(Q \mid z_k = 1)$ es la distribución de probabilidad de las medidas de calidad dado un estado de calidad (k) determinado. Bajo este enfoque, puede requerirse el conocimiento de Δs durante el entreno. Por lo tanto, pueden ser necesarias señales limpias y degradadas (datos estéreo) durante el entreno. Aquí, k puede indicar el k-ésimo componente de un vector binario K-dimensional z_{ik} (los estados de calidad correspondientes para la prueba i). Cada elemento de z_i, z_{ik}, puede representar un estado de calidad. Para cada z_i, uno de sus elementos z_{ik} será 1, y el resto será 0. Entonces, $z_{ik} = 1$ puede corresponder a la indicación de que se considera el estado de calidad de orden k-ésimo. Esto también puede escribirse como z_k. Esta representación puede ser ventajosa en una notación matemática del procedimiento.

De una manera no supervisada Δs puede ser visto en la formación y oculto en pruebas mientras z estar oculto tanto en entreno como en prueba. Esto puede ser ventajoso ya que puede ser menos restrictivo que el enfoque que se ha descrito anteriormente. Sin embargo, todavía se requiere que se proporcionen señales limpias y degradadas (datos estéreo).

De una manera ciega, Δs y z son ocultos en el entreno y las pruebas. Esto puede ser ventajoso para situaciones reales, ya que no se requiere conocer el estado de calidad o la diferencia correspondiente entre las calificaciones vistas y ocultas durante el entreno. Por lo tanto, puede ser particularmente interesante para muchas aplicaciones reales. Pueden extraerse parámetros utilizando un algoritmo adecuado, por ejemplo, un algoritmo de Maximización de Expectativa, por ejemplo, utilizando las expresiones que se dan en el Anexo II.

Alternativamente, puede adaptarse también una Red Bayesiana ya entrenada que pueda haber sido entrenada para un determinado caso, en lugar de formarse desde cero. En tal proceso de adaptación, los parámetros entrenados de la Red Bayesiana pueden adaptarse en función de, por ejemplo, algunos datos (adaptación) disponibles para circunstancias específicas para describir mejor las circunstancias. Dicha Red Bayesiana adaptada puede utilizarse

entonces para estimar la fiabilidad en un procedimiento tal como se ha descrito anteriormente. Esto puede ser particularmente útil si los datos presentes para las circunstancias específicas no serían suficientes para un nuevo entreno de una Red Bayesiana. Sin embargo, estos datos pueden ser suficientes para una adaptación. Por lo tanto, esto puede ser preferible para casos en los que solamente hay presentes pocos datos.

Pueden considerarse varias soluciones para adaptar una Red Bayesiana. Una solución útil puede ser la técnica de Máximo A Posteriori (MAP). Varios parámetros de la BN pueden ser candidatos para ser adaptados, por ejemplo, medias y precisiones que determinan las distribuciones de $P(Q \mid z)$ $P(\Delta s \mid \theta, z)$ y $P(s \mid \theta)$. Pueden encontrarse más explicaciones sobre MAP en el Anexo (III), y en Chin-Hui Lee y Jean-Luc Gauvain, "*MAP Estimation of Continuous Density HMM: Theory and Applications*", *Proceedings of DARPA Speech & Nat. Lang.* 1992.

La Red Bayesiana en un procedimiento tal como se ha descrito anteriormente puede utilizar una, dos, tres, o más medidas de calidad y normalmente la calificación (o resultado) del sistema de reconocimiento de un orador como entrada para entrenar y/o calcular una fiabilidad. Normalmente, estas medidas de calidad se proporcionan a través de uno, dos o más sistemas diferentes.

Utilizando un procedimiento como el que se ha descrito anteriormente, puede calcularse entonces una fiabilidad de la decisión del sistema de reconocimiento de un orador utilizando, por ejemplo, uno, dos, tres, cuatro o más medidas de calidad normalmente calculadas y proporcionadas previamente y normalmente también la calificación (o resultado) del sistema de reconocimiento de un orador.

En un procedimiento como el que se ha descrito anteriormente, la fiabilidad puede utilizarse para tomar una decisión. En base a la fiabilidad, puede tomarse una decisión sobre si la prueba era fiable o no. Esta decisión puede realizarse opcionalmente descartando pruebas fiables basadas en la fiabilidad.

Alternativamente, la decisión puede realizarse opcionalmente transformando un cociente de probabilidad (calibrado) (LR) o un cociente de probabilidad de registro (calibrado) (LLR) o una calificación (calibrada) proporcionada por el sistema de reconocimiento de un orador.

Estos cocientes de probabilidad (calibrados) o cocientes de probabilidad de registro (calibrados) proporcionados por el sistema de reconocimiento de un orador pueden transformarse en función de la fiabilidad. Por ejemplo, el cociente de probabilidad puede pasar a 1 (o si el sistema de reconocimiento de un orador proporciona un registro LR (LLR), el LLR puede pasar a 0) si la probabilidad o fiabilidad de una prueba $P(R_i = \Re \mid \hat{s}_i, Q_i)$ es demasiado baja, por ejemplo, menor que un cierto umbral predeterminado.

Como alternativa, la calificación proporcionada por el sistema de reconocimiento de un orador puede transformarse en un cociente de probabilidad (calibrado) o en un cociente de probabilidad de registro (calibrado) o una calificación (calibrada) utilizando una función que depende de la calificación proporcionada por el sistema de reconocimiento de un orador y la fiabilidad estimada por la Red Bayesiana. Así, en lugar de transformar un cociente de probabilidad (calibrado) proporcionado por el sistema de reconocimiento de un orador, en un sistema de reconocimiento de un orador que proporciona una calificación vista en lugar de un cociente de probabilidad o de probabilidad de registro, la calificación puede transformarse también en un cociente de probabilidad (calibrado), o un cociente de probabilidad de registro (calibrado) o una calificación (calibrada), que depende de la fiabilidad de la Red Bayesiana.

Otra solución para la transformación de los cocientes de probabilidad (calibrados) puede ser un procedimiento que comprende calcular el cociente de probabilidad transformado LR_{trans} (o un cociente de probabilidad de registro transformado o una calificación transformada) como una función del cociente de probabilidad (o cociente de probabilidad de registro o calificación) proporcionado por el sistema de reconocimiento de un orador y la probabilidad posterior (fiabilidad). Esto puede escribirse, por ejemplo, como $LR_{i,trans} = f(LR_i, P(R_i = \Re \mid \hat{s}_i, Q_i))$. La función utilizada para dicha transformación (f) (la función de transformación), puede ser, por ejemplo, una función lineal discriminante definida por un conjunto de parámetros que pueden haber sido entrenados en una fase de desarrollo.

Otra posibilidad para utilizar la fiabilidad calculada por la Red Bayesiana propuesta sería fusionar varios sistemas de reconocimiento de un orador. Si se ha analizado un audio de prueba por varios sistemas de reconocimiento de un orador, las calificaciones de todos ellos pueden fusionarse dando más importancia a los que tienen una mayor fiabilidad. Una posibilidad sería, por ejemplo, ser ponderar las calificaciones por la fiabilidad. Esto puede llevarse a cabo, por ejemplo, mediante la fórmula

$$s_{fused} = \frac{\sum_{j=1}^{J} \hat{s}_j P(R_i = \Re \mid \hat{s}_j, Q_j)}{\sum_{j=1}^{J} P(R_i = \Re \mid \hat{s}_j, Q_j)}$$ $s_{fusionada}$

Aquí, J puede representar el número de sistemas de reconocimiento de un orador. *j* puede estar entre 1 y J, e identificar un sistema de reconocimiento de un orador particular de los J sistemas de reconocimiento de un orador, en el que J puede ser 1, 2, 3 o más. En este caso, se supone que todas las calificaciones de los sistemas de reconocimiento de un orador se encuentran dentro de un mismo rango, por ejemplo, entre 0 y 1. Si ese no es el caso, pueden añadirse términos de compensación adicionales a la expresión mencionada anteriormente para ponerlas en el mismo intervalo, por ejemplo, normalizando todas las calificaciones de reconocimiento de un orador de los diferentes sistemas.

Un procedimiento de acuerdo con la invención tal como se ha descrito anteriormente puede utilizarse para la verificación de un orador y/o para la identificación de un orador puesto que cada prueba se considera de manera independiente.

La invención comprende, además, un medio legible por ordenador que comprende instrucciones legibles por ordenador para la ejecución de un procedimiento como el que se ha descrito anteriormente cuando se ejecuta en un ordenador. La invención también comprende un sistema adaptado para ejecutar un procedimiento tal como se ha descrito anteriormente.

Otros detalles de la invención se explican en las siguientes figuras.

La figura 1 muestra una Red Bayesiana tal como se utiliza en la técnica anterior;
La figura 2 muestra una Red Bayesiana que puede utilizarse para un procedimiento de acuerdo con la invención;
La figura 3 muestra posibles parámetros de entrada a una Red Bayesiana;
Las figuras 4(a), 4(b) y 4(c) muestran diferentes procedimientos de entreno para entrenar una Red Bayesiana
La figura 5 muestra las etapas que pueden utilizarse para la adaptación de una Red Bayesiana;
La figura 6 muestra las etapas de un procedimiento para estimar la fiabilidad de una decisión de un sistema de reconocimiento de un orador;
La figura 7 muestra las etapas que pueden estar comprendidas en un procedimiento de acuerdo con la invención;
La figura 8 muestra una etapa que puede estar comprendida en un procedimiento de acuerdo con la invención.

La figura 1 muestra una Red Bayesiana utilizada por ejemplo en el documento "*A probabilstic measure of modality reliability in speaker verification*", publicado en *Acoustics, Speech and Signal Processing* 2005, *Proceedings*, (ICASSP '05), IEEE *International Conference* en 2005 por J. Richiardi y otros. En el mismo, los nodos vacíos indican variables ocultas, los nodos sombreados indican variables vistas y un pequeño nódulo sólido marca un parámetro determinístico. Un nodo o grupo de nodos rodeados por un cuadro (denominado placa) marcado con N indica que hay N nodos de este tipo, por ejemplo, N pruebas. Los arcos entre nodos apuntan desde las variables padre hacia las variables hijas, que representan las dependencias condicionales entre padres e hijos. Aquí, una variable padre corresponde a una variable de la que depende una variable correspondiente denominada variable hija. Las expresiones utilizadas en la figura 1 son conocidas, por ejemplo, de la referencia de Bishop citada anteriormente.

Las variables utilizadas en la figura 1 son las siguientes. s_i es la calificación de la verificación del orador vista, Q_i representa las medidas de calidad de voz vistas relacionadas con una prueba (sólo SNR en el documento mencionado anteriormente). $\theta_i \in \{T, NT\}$ es la etiqueta oculta de la prueba, donde *T* es la hipótesis de que el entreno y el audio de prueba pertenecen al mismo orador y *NT* es la hipótesis de que el entreno y el audio de prueba pertenecen a diferentes oradores. $\hat{\theta}_i$ es la decisión de reconocimiento de un orador vista para la prueba de orden i-ésima, que se marca con el subíndice i, después de aplicar un umbral ξ_θ. $R_i \in (\Re; U)$ es la fiabilidad oculta de la prueba, donde $\Re$ es la hipótesis de que la decisión es fiable y no fiable. $\pi_\theta = (P_T, P_{NT})$ es la hipótesis determinística previa donde P_T es el objetivo previo y $P_{NT} = 1- P_T$ el no objetivo previo. Finalmente, $\pi_R = (P_\Re, P_U)$ es la fiabilidad determinístico previa. Utilizando la Red Bayesiana, es posible calcular la distribución posterior de R_i dadas las variables vistas y determinísticas $P(R_i \mid S_i, Q_i, \hat{\theta}_i, \pi_\theta, \pi_R)$.

Tal modelo puede presentar el inconveniente de que los parámetros de la Red Bayesiana pueden depender del umbral de verificación del orador ξ_θ. Por lo tanto, una variación del umbral puede hacer que se necesario un re-entreno, que puede no ser una opción o no ser ventajoso en muchos casos reales.

La figura 2 muestra una Red Bayesiana que puede utilizarse en algunas realizaciones de la invención. En la misma, los nodos vacíos indican variables ocultas, los nodos sombreados indican variables vistas y un pequeño nódulo sólido indica un parámetro determinístico. De nuevo, un grupo de nodos rodeados por una caja (llamada placa) marcada con la letra *N* o bien P indica que hay *N* o P grupos de nodos de este tipo. En este ejemplo, puede haber N pruebas y P medidas de calidad, en el que N puede ser 1, 2, 3 o más y en el que P puede ser 1, 2, 3, 4 o más.

Una Red Bayesiana que puede utilizarse para un procedimiento tal como se ha descrito anteriormente puede utilizar o comprender algunos o todos de los siguientes componentes y variables:

$\hat{s}_i$ es la calificación vista proporcionada por el sistema de reconocimiento de un orador. El audio de prueba y el (los) audio(s) modelo puede(n) haber sido degradado(s). En general, tal $\hat{s}_i$ puede ser un vector de calificaciones de diferentes sistemas de reconocimiento de un orador. En otras realizaciones, puede ser una variable escalar. El subíndice *i* que puede ser entre 1 y el número de pruebas *N* ($i \in [1, N]$), puede representar la prueba.

Siempre que una variable tenga un subíndice *i*, significa que se trata de una realización de una variable aleatoria. Tal realización se denomina variable correspondiente, pero tiene el subíndice adicional *i* que no tiene la variable aleatoria. Por ejemplo, $\hat{s}_i$ es una realización de $\hat{s}$.

s_i es la calificación limpia que puede ser un vector de calificaciones limpias, por ejemplo, de diferentes sistemas de reconocimiento de un orador. Puede ser una variable escalar en otras realizaciones. El subíndice *i* que puede ser entre 1 y el número de pruebas *N* ($i \in [1, N]$), puede representar la prueba. Tal calificación limpia correspondería a la calificación proporcionada por un sistema de reconocimiento de un orador sin ninguna degradación de audio de prueba y audio(s) modelo. En un caso general, dicha calificación limpia puede ser una variable oculta. Sin embargo, si el procedimiento comprende un entreno de la Red Bayesiana, dependiendo del entreno, la calificación limpia puede ser vista en la fase de entreno. En particular, si se utiliza una base de datos artificialmente degradada, por ejemplo, añadiendo a las señales ruidos aditivos, o una distorsión convolucional, una calificación limpia puede ser vista en la fase de entreno. La distribución de *s* bajo la condición θ puede suponerse que es gaussiana. $P(s \mid \theta) = N(s; \mu_{s\theta}, \Lambda_{s\theta}^{-1})$ donde θ es la etiqueta de prueba real, que puede ser objetivo o no objetivo ($\theta \in (T, NT)$). $\mu_{s\theta}$, y $\Lambda_{s\theta}^{-1}$ son la media y la varianza (inversa de la precisión) de la distribución (normalmente gaussiana) que siguen las calificaciones limpias asociadas a θ.

Además, la relación entre calificaciones limpias ocultas y vistas pueden ser modeladas siguiendo la expresión $\hat{s}_i = s_i + \Delta s_i$. Δs_i puede ser la desviación (diferencia) entre la calificación vista (con ruido) y la calificación limpia (oculta).

π_θ es la hipótesis previa $\pi_\theta = (P_T, P_{NT})$ con $P_T + P_{NT} = 1$ y puede ser determinístico. Aquí, P_T puede ser el objetivo previo y P_{NT} el no objetivo previo. El objetivo previo es la probabilidad previa de una prueba objetivo. Esto puede considerarse como la probabilidad de una prueba objetivo sin saber nada acerca de una prueba.

z_i son los estados de la calidad (estados de calidad) asociados a la i-ésima prueba. Se trata de un vector binario K-dimensional con elementos z_{ik} con k entre 1 y el número de estados de calidad K ($k \in [1, K]$).

z_i normalmente es un vector binario. Dado z_i normalmente sólo un elemento será igual a 1, mientras que los otros son 0. Puede haber K estados de calidad. Por lo tanto, el elemento z_i que es igual a 1 determina el estado de calidad asociado a la prueba i-ésima, la k-ésima en este caso.

Aunque las medidas de calidad normalmente son variables continuas, la combinación de todas ellas puede discretizarse y afectar a la distribución de Δs_i. La distribución de *z* viene dada por

$$P(z) = \prod_{k=1}^{K} (\pi_{z_k})^{z_k}$$

.

π_z son los coeficientes de la distribución opcionalmente discreta que describe *z*. π_z normalmente es un vector K-dimensional con elementos π_{zk}, en el que π_{zk} normalmente es la probabilidad del estado de calidad k-ésimo (que normalmente es la probabilidad de z_k).

π_z puede ser una variable de la Red Bayesiana y normalmente se obtiene durante la fase de entreno de la Red Bayesiana. También puede haber otras variables de la Red Bayesiana que sean entrenados durante la fase de entreno. La *z* K-dimensional determina un estado de calidad. Cuando está asociada a una prueba, normalmente se denomina z_i.

Así, la probabilidad de *z* normalmente es π_{zk}, en el que z_k es el elemento de *z* que es 1. Esto puede expresarse, por ejemplo, tal como se ha dado anteriormente

$$P(z) = \prod_{k=1}^{K} (\pi_{z_k})^{z_k}$$

Q_{pi} son las medidas de calidad vistas. Se considera que hay P grupos de medidas de calidad que son independientes entre sí dado z_i ($p \in [1, P]$). Esto puede permitir forzar la independencia entre variables, por ejemplo, variables que no debe ser correlacionadas. Aquí, i puede ser el número de la prueba, y p puede variar entre 1 y el número de medidas de calidad P. Si Q_p se modela por gaussianos esto puede ser lo mismo que tener una matriz de covarianza diagonal por bloques Gaussiana. Aquí, Q_p describe las medidas de calidad vistas. Cuando se refieren a una prueba particular, se hace referencia a Q_{pi}.

Este conjunto puede indicarse como $Q_i = \{Q_{pi}\}_{p=1}^{P}$.

μ_{Qp} y Λ_{Qp} son la media y la precisión (descritas normalmente mediante una matriz) de las distribuciones normalmente gaussianas que describen Q_p. Hay K distribuciones diferentes, tanto como estados de calidad de manera que

$$P(Q \mid z_k = 1) = \prod_{p=1}^{P} N\left(Q_p; \mu_{Q_{pk}}, \Lambda_{Q_{pk}}^{-1}\right)$$

$\mu_{\Delta s}$ y $\Lambda_{\Delta s}$ son la media $\mu_{\Delta s}$ y la precisión $\Lambda_{\Delta s}$ (descritas normalmente mediante una matriz) de la distribución normalmente gaussiana que describe Δs. Hay *2K* distribuciones diferentes, una para cada estado de calidad y θ.

Por lo tanto, $P(\hat{s} \mid s, z_k = 1, \theta) = N(\hat{s}; s + \mu_{\Delta s_{k\theta}}, \Lambda_{\Delta s_{k\theta}}^{-1})$.

Aquí, puede haber N grupos de nodos que comprenden las variables $\hat{s}_i$, s_i, Q_{pi}, z_i y θ_i (un grupo por cada $i \in [1, N]$) y P grupos de nodos que comprenden las variables Q_{pi}, μ_{Qp} y Λ_{Qp} (un grupo por cada $i \in [1, P]$). En particular $\hat{s}_i$ puede depender de z_i, s_i, θ_i, $\mu_{\Delta s}$ y $\Lambda_{\Delta s}$. θ_i puede depender de la π_θ (opcionalmente determinística), s_i puede depender de θ_i, Λ_s, μ_s mientras que μ_s puede depender de Λ_s. z, puede depender de π_z, $\mu_{\Delta s}$ puede depender de $\Lambda_{\Delta s}$, Q_{pi} puede ser una variable vista que dependa de z_i, μ_{Qp} y Λ_{Qp}, mientras que μ_{Qp} puede depender de Λ_{Qp}. $\hat{s}_i$ y Q_{pi} pueden ser vistas, $\Lambda_{\Delta s}$, $\mu_{\Delta s}$, Λ_s, μ_s, θ_i, S_i, Λ_{Qp}, π_z, z_i, y μ_{Qp} pueden ser variables ocultas y π_θ puede ser determinístico.

En la Red Bayesiana que se ha descrito anteriormente, p normalmente adoptará valores entre 1 y P, e i normalmente adoptará valores entre 1 y N.

Aquí, P es el número de medidas de calidad y N el número de pruebas.

La figura 3 incluye un diagrama que muestra parámetros de entrada y de salida de la Red Bayesiana. En particular, la Red Bayesiana utiliza la calificación de un sistema de reconocimiento de un orador y las medidas de calidad elegidas, por ejemplo, para calcular (estimar) la fiabilidad. En otras realizaciones, estos parámetros de entrada pueden utilizarse para el entreno de la Red Bayesiana. En particular, en este caso representado, el parámetro de calidad, la relación señal ruido, el índice de modulación, la entropía y la probabilidad de registro de modelo de fondo universal se mencionan explícitamente. En la figura se indica que pueden utilizarse adicionalmente otros parámetros de calidad.

En otras realizaciones, pueden utilizarse sólo una, dos, tres o cuatro de las medidas de calidad mencionadas o puede utilizarse cualquier número de medidas de calidad que se muestran en combinación con cualquier otra medida de calidad que no se muestra aquí.

Como resultado, puede estimarse (calcularse) la fiabilidad $P(R_i = \Re \mid \hat{s}_i, Q_i)$ del resultado del sistema de reconocimiento, normalmente para un audio de prueba particular y audio(s) modelo particular(es). El resultado puede ser, por ejemplo, que la probabilidad de la decisión de una prueba que se ha encontrado, por ejemplo, comparando la calificación vista calculada por el sistema de reconocimiento de un orador con un umbral es fiable.

Para calcular la fiabilidad, adicionalmente el umbral de reconocimiento de un orador utilizado por el sistema de reconocimiento de un orador y/o un umbral de fiabilidad normalmente tienen que proporcionarse como parámetros de entrada para la Red Bayesiana también (no mostrado).

La figura 4 muestra tres procedimientos de entreno diferentes que pueden ser utilizados para entrenar la Red Bayesiana.

En particular, la Red Bayesiana puede ser entrenada utilizando datos de desarrollo estéreo (datos donde hay presentes datos degradados y limpios) en un entreno supervisado. En éste, Δs y z son vistas durante el entreno. Los parámetros se extraen por maximización de expectación o cualquier otro algoritmo adecuado (figura 4 (a)).

La figura 4 (b) muestra un enfoque de entrenamiento diferente para una Red Bayesiana. En el mismo, se utilizan datos de desarrollo estéreo (que comprende datos limpios y datos degradados) en un entrenamiento sin supervisión. En tal entreno, Δs puede ser vista durante el entreno mientras z puede ser oculta durante el entrenamiento. De nuevo, los parámetros del modelo pueden extraerse utilizando un algoritmo adecuado tal como, por ejemplo, un algoritmo de maximización de expectación.

La figura 4 (c) muestra el entrenamiento ciego de la Red Bayesiana. En particular, puede no ser necesario proporcionar datos estéreo. Los datos utilizados para el entrenamiento de la Red Bayesiana en el entrenamiento ciego normalmente están degradados. Cualquier degradación que no se aprecia en las señales de voz de desarrollo normalmente no se modelará mediante la Red Bayesiana. Esto normalmente también es cierto para otros procedimientos de entrenamiento, por ejemplo, tal como se describe respecto a las figuras 4a y 4b. Por lo general, la exactitud de la Red Bayesiana depende de la falta de coincidencia entre los datos desarrollados utilizados para entrenar la Red Bayesiana y los datos de prueba. Con una baja falta de coincidencia, la exactitud de la Red Bayesiana será alta, y viceversa.

En el entrenamiento ciego, Δs y z son variables ocultas en el entrenamiento. Los parámetros se extraen mediante un algoritmo adecuado tal como, por ejemplo, un algoritmo de maximización de expectación.

La figura 5 muestra las etapas que pueden utilizarse en un procedimiento de acuerdo con la invención para la adaptación de una Red Bayesiana (sus parámetros). Partiendo de los datos de adaptación y utilizando los parámetros de una Red Bayesiana que ya ha sido entrenada, puede adaptarse la Red Bayesiana (sus parámetros). Los datos de adaptación pueden comprender la(s) calificación(es) vista(s) ($\hat{s}_i$) proporcionas por el sistema de reconocimiento de un orador partir de los datos de adaptación y una, dos, tres o más medidas de calidad de audios utilizados para la adaptación. Normalmente, los datos de adaptación comprenden todas las medidas de calidad derivadas del uno o más audio(s) utilizado(s) para la adaptación que se consideran en la Red Bayesiana y la(s) calificación(es) vista(s) proporcionada(s) por el sistema de reconocimiento de un orador. Por lo general, la(s) medida(s) de calidad y o la(s) calificación(es) no se calcula(n) a partir del (de los) audio(s) en el entrenamiento de adaptación, sino, por ejemplo, antes del entrenamiento de adaptación. Durante el entrenamiento de adaptación Δs y z pueden ser ocultas.

Tal adaptación puede realizarse, por ejemplo, utilizando un algoritmo de máximo a posteriori (MAP).

Con este enfoque, tras la adaptación, puede estar presente un conjunto de parámetros adaptado de la Red Bayesiana. Por lo tanto, la Red Bayesiana puede utilizarse entonces con los parámetros adaptados.

Tal proceso de adaptación puede ser particularmente útil, aunque solamente esté presente un pequeño conjunto audios modelo en la situación para la cual el modelo debe entrenarse. A continuación, el resultado que puede conseguirse utilizando una Red Bayesiana ya entrenada y adaptando sus parámetros/ adaptando la Red Bayesiana, puede ser más fiable que iniciar el proceso de entrenamiento con la cantidad (limitada) de datos disponibles para la situación particular desde cero.

La figura 6 muestra las etapas de una realización del procedimiento de la invención. En particular, utilizando medidas de calidad de los audios de prueba y modelo y la calificación de un sistema de reconocimiento de un orador que puedan derivarse directa o indirectamente de las pruebas y audios modelo, la Red Bayesiana trabaja con parámetros entrenados que pueden calcular la fiabilidad y tomar una decisión basada en esa fiabilidad. Normalmente es necesario un umbral de reconocimiento de un "orador" y/o un umbral de fiabilidad para tomar una decisión final.

Tal como se ha explicado anteriormente, tal decisión puede ser, por ejemplo, un descarte de una prueba si la decisión no es fiable, una transformación de la calificación, por ejemplo, utilizando una de las funciones descritas anteriormente para ese fin o la fusión de varios sistemas (todos estos no se muestran en la figura 6).

Por ejemplo, una calificación obtenida por el sistema de reconocimiento de un orador puede ser transformada en un cociente de probabilidad transformado o un cociente de probabilidad de registro transformado dependiente de la fiabilidad para obtener un cociente de probabilidad (registro) transformado (calibrado). Por lo tanto, a partir de un sistema de reconocimiento de un orador que proporciona una calificación bruta que no se da como un cociente de probabilidad (registro), la calificación puede ser transformada en un cociente de probabilidad transformado (calibrado) o un cociente de probabilidad de registro transformado (calibrado) LLR o una calificación transformada (calibrada) en un formato diferente de un cociente de probabilidad (registro), o a partir de un sistema de

reconocimiento de un orador que proporcione un cociente de probabilidad (calibrado) o un cociente de probabilidad de registro (calibrado) la LR (LLR) puede transformarse en vista de las fiabilidades estimadas por la Red Bayesiana para resultar en un LLR transformado (calibrado) o un LR transformado (calibrado) (no mostrado en la figura 6).

La figura 7 muestra cómo puede calcularse una calificación final utilizando la fiabilidad de las calificaciones y las calificaciones de los diversos sistemas de reconocimiento de un orador 1 a M (donde M es el número de sistemas de reconocimiento de un orador diferentes y puede ser 1, 2, 3, 4 o más) en un diagrama. Esta calificación final puede corresponder a una decisión mencionada, por ejemplo, en la figura 6. En particular, partiendo de los datos que suelen ser un audio de prueba y audio(s) modelo, varios sistemas de reconocimiento de un orador, en este caso, 1 a M, calculan la calificación de 1 a M. Aquí, cada sistema de reconocimiento de un orador proporciona entonces su calificación a la Red Bayesiana. Utilizando las medidas de calidad de audio de prueba y audio(s) modelo y la calificación de los sistemas de reconocimiento de un orador, la Red Bayesiana procede entonces a tomar una decisión. Las medidas de calidad normalmente se extraen de los datos mediante un módulo externo. Este módulo, sin embargo, puede estar integrado también con la Red Bayesiana en otras realizaciones. La decisión puede ser, por ejemplo, una calificación final que pueda considerarse contra el umbral. Para tomar tal decisión, puede utilizarse otra Red Bayesiana.

En otras realizaciones, algún otro módulo diferente de la Red Bayesiana puede tomar la decisión utilizando la entrada de la Red Bayesiana. Por ejemplo, las calificaciones pueden fusionarse mediante un módulo externo de acuerdo con su fiabilidad que puede obtenerse con la Red Bayesiana explicada.

En particular, una calificación final puede ser una combinación de calificaciones ponderadas en el que las calificaciones con mayor fiabilidad pesen más que las calificaciones con menor fiabilidad.

En particular, en dicha fusión, una Red Bayesiana puede calcular la fiabilidad de las pruebas proporcionadas por todos sistemas de reconocimiento de un orador, o pueden utilizarse dos, tres, o más Redes Bayesianas. En particular, para cada calificación de un sistema de reconocimiento de un orador, puede utilizarse una Red Bayesiana para calcular la fiabilidad y puede tomarse entonces la decisión en la etapa siguiente (no mostrado en la figura 7). Normalmente, cuando se cambia el sistema de identificación de un orador, la Red Bayesiana tiene que entrenarse de nuevo. Por lo tanto, en algunas realizaciones, pueden utilizarse dos, tres, o más Redes Bayesianas. En otras realizaciones puede utilizarse solamente una Red Bayesiana.

La figura 8 muestra también una etapa que puede estar comprendida en un procedimiento de acuerdo con la invención. Una Red Bayesiana puede utilizar las medidas de calidad de entrada y el resultado de un sistema de reconocimiento de un orador, por ejemplo, una calificación vista $\hat{s}_i$, por ejemplo, un cociente de probabilidad de registro (calibrado) (LLR) o un cociente de probabilidad (calibrado) (LR) como entrada. Entonces puede calcularse la fiabilidad del resultado del sistema de reconocimiento de un orador.

En base a la fiabilidad que se calcula, puede tomarse entonces una decisión. Esto puede realizarse, por ejemplo, mediante el cálculo de un cociente de probabilidad transformado (calibrado) o un cociente de probabilidad de registro transformado (calibrado) o una calificación transformada (calibrada), en base a la fiabilidad y el resultado del sistema de reconocimiento de un orador. Sin embargo, normalmente, si el resultado del sistema de reconocimiento de un orador es un cociente de probabilidad o un cociente de probabilidad de registro, no puede calcularse ninguna calificación transformada en un formato diferente de un cociente de probabilidad (registro).

Si un cociente de probabilidad o cociente de probabilidad de registro es el resultado de un sistema de reconocimiento de un orador, puede calcularse un cociente de probabilidad transformado (calibrado) o un cociente de probabilidad transformado de registro (calibrado) como salida, utilizando la fiabilidad.

A partir de un cociente de probabilidad como resultado de un sistema de reconocimiento de un orador, puede calcularse un cociente de probabilidad transformado (calibrado) o un cociente de probabilidad transformado de registro (calibrado). En consecuencia, a partir de un cociente de probabilidad de registro puede calcularse como resultado un cociente de probabilidad transformado (calibrado) o un cociente de probabilidad de registro transformado.

Alternativamente, puede calcularse una calificación transformada (calibrada) en un formato diferente al de un cociente de probabilidad (registro) utilizando $\hat{s}_i$ en un formato diferente que una relación de un cociente de probabilidad (registro).

El cociente de probabilidad transformado y/o el cociente de probabilidad de registro transformado o la calificación transformada pueden estar ser calibrados o no. El cociente de probabilidad de registro el cociente de probabilidad o la calificación proporcionada por un sistema de reconocimiento de un orador también pueden ser calibrados o pueden no ser calibrados.

Las etapas de calcular una decisión (por ejemplo, un cociente de probabilidad transformado o un cociente de probabilidad de registro transformado o una calificación transformada) basado en el resultado del sistema de reconocimiento de un orador (que puede ser, por ejemplo, una calificación $\hat{s}_i$ en un formato diferente de un cociente de probabilidad (registro) o un cociente de probabilidad de registro o un cociente de probabilidad), utilizando la fiabilidad estimada por la Red Bayesiana puede realizarse por un módulo o sistema diferente de una Red Bayesiana, en el que la fiabilidad puede ser proporcionada por la Red Bayesiana y el resultado del sistema de reconocimiento de un orador puede ser proporcionado por el sistema de reconocimiento de un orador como entrada para el módulo o sistema.

ANEXO I

La probabilidad posterior de la calificación oculta, dada la calificación vista y las medidas de cantidad, *P (s | ŝ, Q*) puede expresarse como (puede encontrarse también un procedimiento para calcular la probabilidad posterior de la calificación oculta dado el resultado visto y las medidas cuantitativas, por ejemplo, en J. Villalba: *A Bayesian Network for Reliability Estimation: Unveiling the Score Hidden under the Noise*, Informe Técnico, Universidad de Zaragoza, Zaragoza (España), 2012):

$$P(s \mid \hat{s}, Q) = \sum_{\theta \in \{T, NT\}} \sum_{k=1}^{K} P(s, \theta, z_k = 1 \mid \hat{s}, Q) = \sum_{\theta \in \{T, NT\}} \sum_{k=1}^{K} P(s \mid \hat{s}, Q, \theta, z_k = 1) P(\theta, z_k = 1 \mid \hat{s}, Q)$$

donde puede demostrarse que *P (s | ŝ, Q, θ, z_k = 1*) sigue una distribución Gaussiana.

N (s; $\mu_{sk\theta}$, $\Lambda_{sk\theta}^{-1}$), donde la media y la precisión son, respectivamente:

$\Lambda_{sk\theta}^{-1} = \Lambda_{sk\theta} + \Lambda_{s\theta}$

$\mu_{sk\theta} = \Lambda_{sk\theta}^{-1} (\Lambda_{\Delta sk\theta} (\hat{s} - \mu_{\Delta\Delta sk\theta}) + \Lambda_{s\theta} + \mu_{s\theta}$

Por otra parte, utilizando la regla de Bayes,

$$P(\theta, z_k = 1 \mid \hat{s}, \theta) = \frac{P(\hat{s} \mid Q, z_k = 1) P(Q \mid z_k = 1) P(\theta) \pi_{z_k}}{\sum_{\theta \in \{T, NT\}} \sum_{k=1}^{K} P(\hat{s} \mid \theta, z_k = 1) P(Q \mid z_k = 1) P(\theta) \pi_{z_k}}$$

donde

$$P(Q \mid z_k = 1) = \prod_{p=1}^{P} N(Q_p; \mu_{Q_{pk}}, \Lambda_{Q_{pk}}^{-1})$$

$$P(\hat{s} \mid \theta, z_k = 1) = N(\hat{s}; \mu'_{\hat{s}_{k\theta}}, \Lambda'^{-1}_{\hat{s}_{k\theta}})$$

$$\mu'_{\hat{s}_{k\theta}} = \mu_{s_\theta} + \mu_{\Delta s_{k\theta}}$$

$$\Lambda'^{-1}_{\hat{s}_{k\theta}} = \Lambda_{s_\theta} \Lambda'^{-1}_{s_{k\theta}} \Lambda_{\Delta s_{k\theta}}$$

ANEXO II

El algoritmo EM es un procedimiento iterativo que estima los parámetros de un modelo estadístico que tiene algunas variables latentes mediante el uso de probabilidad máxima como objetivo. La iteración EM alterna entre realizar una etapa de expectativa (E), que crea una función para la expectativa del registro-probabilidad evaluado utilizando la estimación actual para los parámetros, y una etapa de maximización (M), que calcula parámetros maximizando el registro-probabilidad esperado encontrado en la etapa E. Estas estimaciones de los parámetros se utilizan entonces para determinar la distribución de las variables latentes en la siguiente etapa E. (Un procedimiento de utilización de un algoritmo EM para extraer los parámetros de un modelo estadístico puede encontrarse también, por ejemplo, en J. Villalba: *A Bayesian Network for Reliability Estimation: Unveiling the Score Hidden under the Noise*, Informe Técnico, Universidad de Zaragoza, Zaragoza (España), 2012).

Etapa E

Es la variable definida $\gamma(z_k) = P(z_k = 1 \mid \hat{s}, Q, \theta)$ que puede calcularse como:

$$\gamma(z_k) = \frac{\pi_{z_k} P(\hat{s} \mid z_k = 1, \theta) P(Q \mid z_k = 1)}{\sum_{k=1}^{K} \pi_{z_k} P(\hat{s} \mid z_k = 1, \theta) P(Q \mid z_k = 1)}$$

$$P(Q \mid z_k = 1) = \prod_{p=1}^{P} N(Q_p; \mu_{Q\,pk}, \Lambda_{Q_{pk}}^{-1})$$

$$P(\hat{s} \mid \theta, z_k = 1) = N(\hat{s}; \mu_{\hat{s}_{k\theta}}^{'}, \Lambda_{\hat{s}_{k\theta}}^{'-1})$$

$$\mu_{\hat{s}_{k\theta}}^{'} = \mu_{s_\theta} + \mu_{\Delta s_{k\theta}}$$

$$\Lambda_{\hat{s}_{k\theta}}^{'} = \Lambda_{s_\theta} \Lambda_{s_{k\theta}}^{'-1} \Lambda_{\Delta s_{k\theta}}$$

Etapa M

La etapa M proporciona la nueva estimación de los parámetros modelo una vez que se ha llevado a cabo la etapa E:

$$\pi_{z_k} = \frac{\sum_{i=1}^{N} \gamma(z_{ik})}{\sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_{ik})}$$

$$\mu_{Q\,pk} = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) Q_{pi}}{\sum_{i=1}^{N} \gamma(z_{ik})}$$

$$\Lambda_{Q_{pk}}^{-1} = \frac{\sum_{i=1}^{N} \gamma(z_{ik})(Q_{pi} - \mu_{Q_{pk}})(Q_{pi} - \mu_{Q_{pk}})^T}{\sum_{i=1}^{N} \gamma(z_{ik})}$$

$$\mu_{s_\theta} = \frac{\sum_{i=1}^{N} t_{i\theta} E[s_i]}{\sum_{i=1}^{N} t_{i\theta}}$$

$$\Lambda_{s_\theta}^{-1} = \frac{\sum_{i=1}^{N} t_{i\theta} E[s_i s_i^T]}{\sum_{i=1}^{N} t_{i\theta}} - \mu_{s_\theta} \mu_{s_\theta}^T$$

$$\mu_{\Delta s_{k\theta}} = \frac{\sum_{i=1}^{N} t_{i\theta} \gamma(z_{ik})(\hat{s}_i - \mu_{si_{k\theta}}')}{\sum_{i=1}^{N} t_{i\theta} \gamma(z_{ik})}$$

$$\Lambda_{\Delta s_{k\theta}}^{-1} = \frac{\sum_{i=1}^{N} t_{i\theta} \gamma(z_{ik})(\hat{s}_i - \mu_{si_{k\theta}}')(\hat{s}_i - \mu_{si_{k\theta}}')^T}{\sum_{i=1}^{N} t_{i\theta} \gamma(z_{ik})} + \Lambda_{S_{k\theta}}'^{-1} - \mu_{\Delta s_{k\theta}} \mu_{\Delta s_{k\theta}}^T$$

$$\mu_{si_{k\theta}}' = \Lambda_{s_{k\theta}}'^{-1} (\Lambda_{\Delta s_{k\theta}} (\hat{s}_i - \mu_{\Delta s_{k\theta}}) + \Lambda_{s_\theta} \mu_{s_\theta}$$

$$\Lambda_{s_{k\theta}}' = \Lambda_{\Delta s_{k\theta}} + \Lambda_{s_\theta}$$

donde $t_{i\theta} = 1$ si $\theta_i = \theta$, y $t_{i\theta} = 0$ si $\theta_i \neq \theta$. *E* es el operador de expectación.

ANEXO III

Se utiliza un algoritmo de máximo a posteriori A para adaptar los medios y covarianzas de P (Q | z), y P (Δs | θ, z) y P (s | θ) con unos pocos datos de destino. Teniendo en cuenta los medios y covarianzas correspondientes inicialmente incluidos en la Red Bayesiana ($\mu_{0,Q_{pk}}$, $\sum_{0,Q_{pk}}$, $\mu_{0,\Delta s_{k\theta}}$, $\sum_{0,\Delta s_{k\theta}}$, $\mu_{0,s_{k\theta}}$ y $\sum_{0,s_{k\theta}}$), que se han obtenido con los datos de desarrollo; y los medios y covarianzas extraídos por el procedimiento de entreno de la Red Bayesiana con los datos objetivo (véase en Anexo II, $\mu_{ML,Q_{pk}}$, $\sum_{ML,Q_{pk}}$, $\mu_{ML,\Delta s_{k\theta}}$, $\sum_{ML,\Delta s_{k\theta}}$, $\mu_{ML,s_{k\theta}}$ y $\sum_{ML,s_{k\theta}}$), los parámetros adaptados se obtienen por regresión lineal según la cantidad de datos de destino:

$$\mu_{Q_{pk}} = \frac{1}{\beta_k}(\beta_0 \mu_{0,Q_{pk}} + N_k \mu_{ML,Q_{pk}})$$

$$\sum_{Q_{pk}} = \frac{1}{\rho_k}\left((\rho_0 \sum_{0,Q_{pk}} + N_k \sum_{ML,Q_{pk}} + \frac{\beta_0 N_k}{\beta_k}(\mu_{ML,Q_{pk}} - \mu_{0,Q_{pk}})(\mu_{ML,Q_{pk}} - \mu_{0,Q_{pk}})^T\right)$$

$$\mu_{\Delta s_{k\theta}} = \frac{1}{\beta_k}\left(\beta_0 \mu_{0,\Delta s_{k\theta}} + N_k \mu_{ML,\Delta s_{k\theta}}\right)$$

$$\sum_{\Delta s_{k\theta}} = \frac{1}{\rho_k}\left((\rho_0 \sum_{0,\Delta s_{k\theta}} + N_k \sum_{ML,\Delta s_{k\theta}} + \frac{\beta_0 N_k}{\beta_k}(\mu_{ML,\Delta s_{k\theta}} - \mu_{0,\Delta s_{k\theta}})(\mu_{ML,\Delta s_{k\theta}} - \mu_{0,\Delta s_{k\theta}})^T\right)$$

$$\mu_{s_{k\theta}} = \frac{1}{\beta_k}(\beta_0 \mu_{0,s_{k\theta}} + N_k \mu_{ML,s_{k\theta}})$$

$$\sum_{s_{k\theta}} = \frac{1}{\rho_k}\left((\rho_0 \sum_{0,s_{k\theta}} + N_k \sum_{ML,s_{k\theta}} + \frac{\beta_0 N_k}{\beta_k}(\mu_{ML,s_{k\theta}} - \mu_{0,s_{k\theta}})(\mu_{ML,s_{k\theta}} - \mu_{0,s_{k\theta}})^T\right)$$

Donde β_0, ρ_0 son los factores relevantes para los medios y covarianzas, y N_k es el número de pruebas que pertenecen a un estado de calidad k en los datos de destino. También,

$$\beta_k = N_k + \beta_0$$

$$\rho_k = N_k + \rho_0$$

REIVINDICACIONES

1- Procedimiento para estimar la fiabilidad de un resultado de un sistema de reconocimiento de un orador respecto a un audio de prueba o una impresión de voz de prueba y un modelo de orador, que está basado en un audio modelo, utilizando el procedimiento una Red Bayesiana para estimar si el resultado es fiable, en el que, para estimar la fiabilidad del resultado del sistema de reconocimiento de un orador, se utiliza una, dos, tres, cuatro, o más de cuatro medidas de la calidad del audio de prueba y una, dos, tres, cuatro, o más de cuatro medidas de calidad del modelo de audio,

en el que la Red Bayesiana utiliza

como nodos que describen parámetros vistos en una calificación vista $\hat{s}_i$ y las medidas de calidad Q_i, en el que un índice *i* indica una prueba i-ésima, como nodos que describen parámetros ocultos como una calificación oculta s_i, estados se calidad z_i, coeficientes de la distribución que describen los estados de calidad, media y precisión que describen los grupos de las medidas de calidad, media y precisión que describen la distribución de la desviación Δs_i entre la calificación vista y oculta, media y precisión que describen la distribución de la calificación oculta; y la etiqueta real de la prueba y como nodo que describe un valor determinístico una hipótesis previa, en el que

la calificación vista depende de los estados de calidad, la calificación limpia, la etiqueta de prueba real y la media y la precisión de la distribución que describe la desviación entre la calificación vista y la calificación oculta, y

la etiqueta de prueba real depende de la hipótesis previa,

la calificación oculta depende de la etiqueta real de la prueba y la media y la precisión de la distribución que describe la calificación limpia,

los estados de calidad dependen de los coeficientes de la distribución opcionalmente discreta que describe los estados de calidad,

las medidas de calidad vistas dependen de los estados de calidad y la media y la precisión de la distribución que describe los grupos de medidas de calidad vista,

en el que la probabilidad de fiabilidad $P(R_i = \Re \mid \hat{s}_i, Q_i)$ se calcula utilizando la siguiente expresión:

$$P(R_i = \Re \mid \hat{s}_i, Q_i) = \begin{cases} P(s_i \geq \varphi_\theta \mid \hat{s}_i, Q_i) = \int_{\varphi_\theta}^{\infty} P(s_i \mid \hat{s}_i, Q_i) ds_i \quad if \quad \hat{\theta}_i = T \\ P(s_i < \varphi_\theta \mid \hat{s}_i, Q_i) = \int_{-\infty}^{\varphi_\theta} P(s_i \mid \hat{s}_i, Q_i) ds_i \quad if \quad \hat{\theta}_i = NT \end{cases}$$

donde $R_i \in (\Re; U)$ es la fiabilidad oculta de la prueba, donde $\Re$ es la hipótesis de que la decisión es fiable y U no fiable,

en el que $\hat{\theta}_i = T$ es la decisión del sistema de reconocimiento de un orador de que la persona es el objetivo,

en el que $\hat{\theta}_i = NT$ es la decisión del sistema de reconocimiento de un orador de que la persona no es el objetivo, y en el que φ_θ es el umbral del sistema de reconocimiento de un orador.

2. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que la media que describe la desviación entre la calificación vista y limpia depende de la precisión que describe la desviación entre la calificación vista y limpia y
en el que la media que describe las medidas de calidad opcionalmente depende de la precisión que describe las medidas de calidad.

3. Procedimiento de acuerdo con una de las reivindicaciones anteriores, que comprende, además, entrenar la Red Bayesiana antes de que se utilice para estimar la fiabilidad de un resultado del sistema de reconocimiento de un orador.

4. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que para el entreno de la Red Bayesiana se utiliza una, dos, tres, cuatro, o más de cuatro medidas de calidad.

5. Procedimiento de acuerdo con una de las reivindicaciones 1 - 4, en el que las medidas de calidad comprenden una, dos, tres o cuatro de las siguientes:

a) relación señal ruido
b) índice de modulación
c) entropía
d) probabilidad de registro de modelo de fondo universal.

6. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que la Red Bayesiana se entrena utilizando un algoritmo de Maximización de Expectación para extraer los parámetros del modelo.

7. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que la Red Bayesiana se entrena de una de las siguientes maneras:

a) supervisada, en la que Δs y z son vistas en el entreno y ocultas en la prueba
b) no supervisada, en la que Δs es vista en el entreno y oculta en la prueba, mientras que z es oculta tanto en el entreno como en la prueba
c) ciega, en la que Δs y z son ocultas en el entreno y en la prueba.

8. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que la Red Bayesiana está adaptada con el fin de describir mejor determinadas circunstancias.

9. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que las medidas de calidad son proporcionadas por uno, dos o más sistemas diferentes de la Red Bayesiana.

10. Procedimiento de acuerdo con una de las reivindicaciones anteriores, caracterizado por el hecho de que la fiabilidad se utiliza para tomar una decisión, que comprende opcionalmente una de las siguientes:

a) descartar pruebas no fiables
b) transformar una calificación;
c) fusionar los resultados de dos, tres o más sistemas de reconocimiento de un orador.

11. Procedimiento de acuerdo con una de las reivindicaciones anteriores, en el que el sistema de reconocimiento de un orador se utiliza para la verificación de un orador y/o la identificación de un orador.

12. Medio legible por ordenador, que comprende instrucciones legibles por ordenador para ejecutar un procedimiento de acuerdo con una reivindicación de acuerdo con una de las reivindicaciones 1 -11 cuando se ejecuta en un ordenador.

13. Sistema adaptado para ejecutar un procedimiento de acuerdo con una de las reivindicaciones 1 -11.

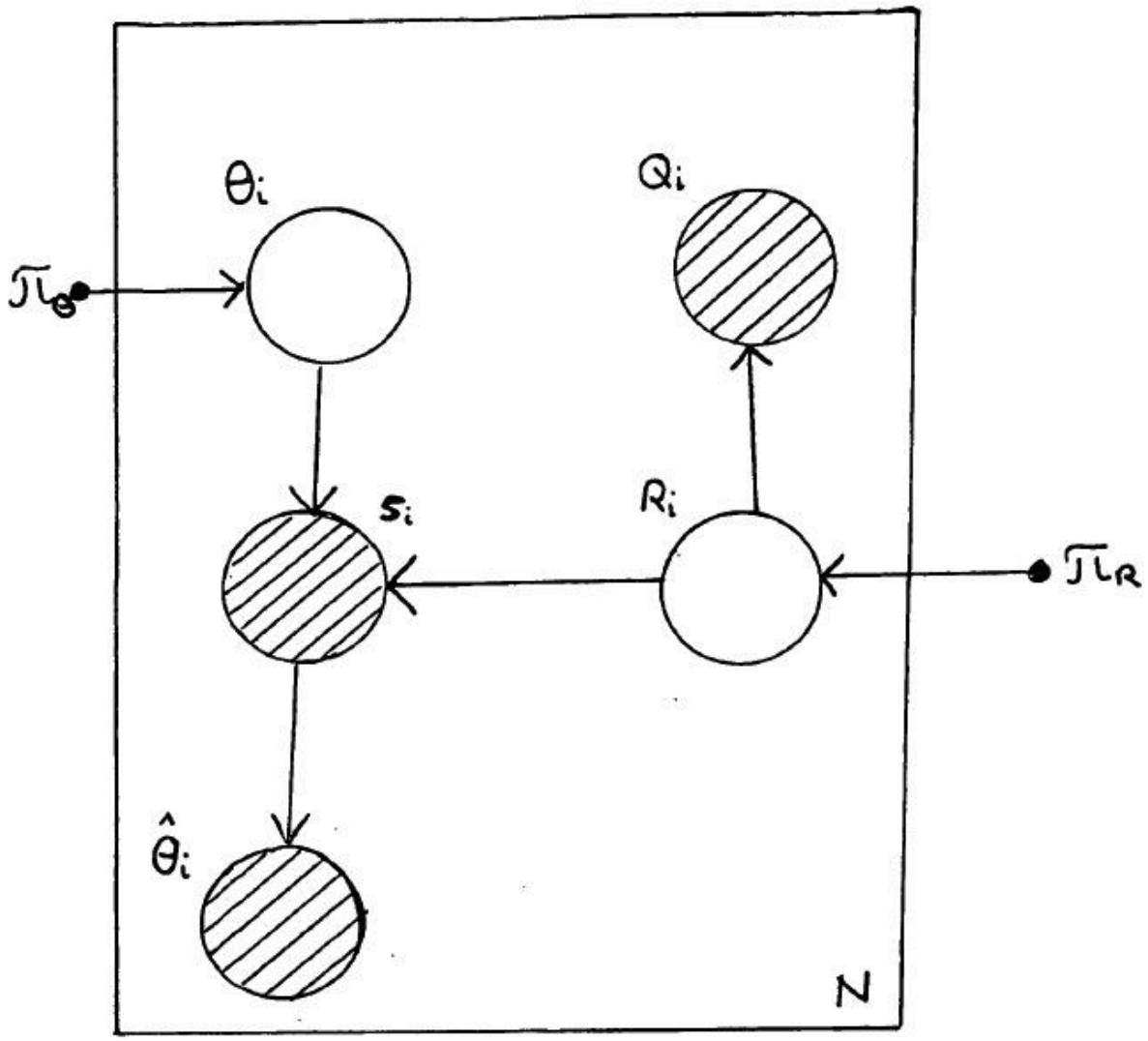

Fig. 1

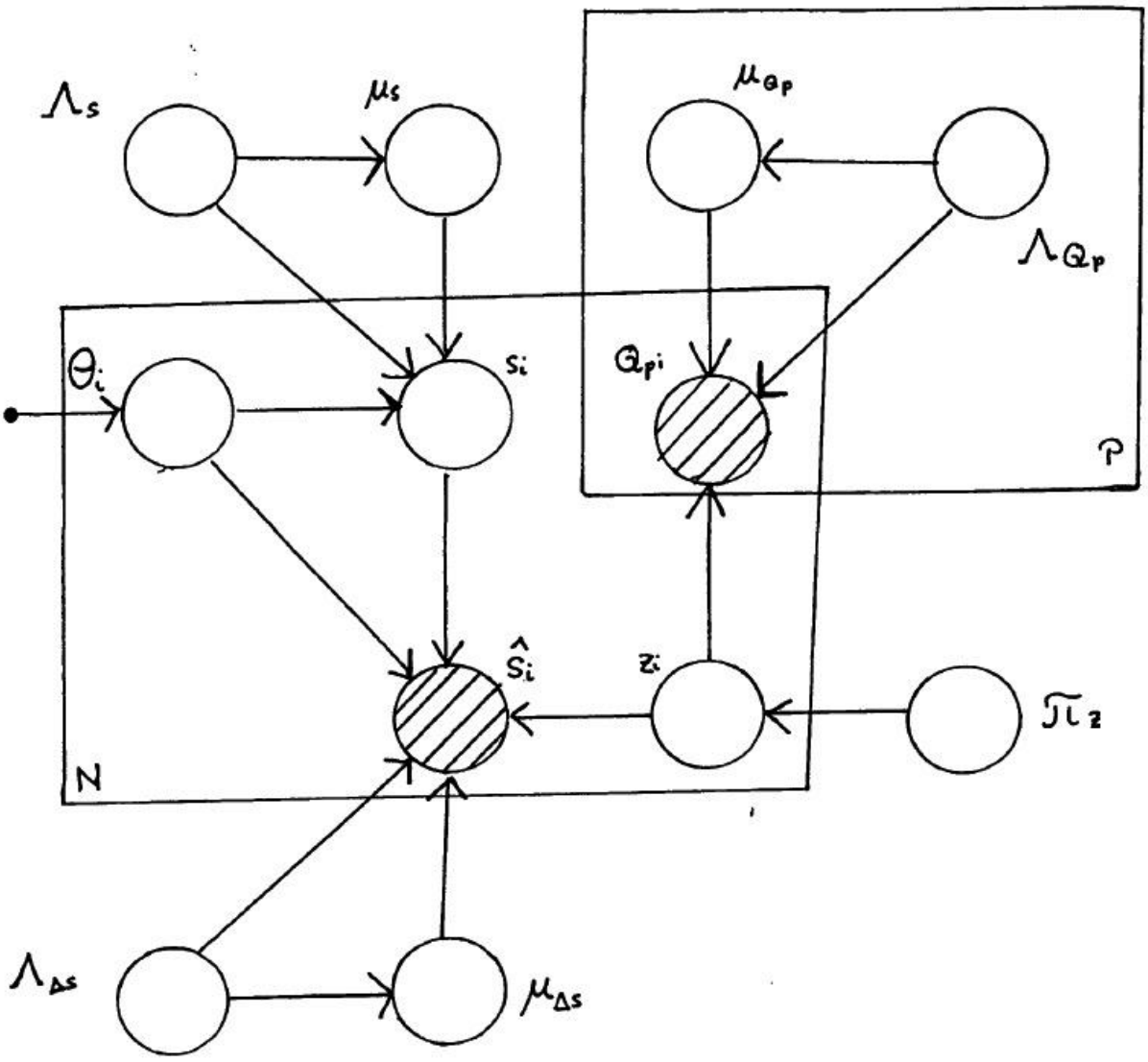
Λ_s
μ_s
μ_{Qp}
Λ_{Qp}
θ_i
s_i
Q_{pi}
P
$\hat{s}_i$
z_i
π_z
N
$\Lambda_{\Delta s}$
$\mu_{\Delta s}$

fig. 2

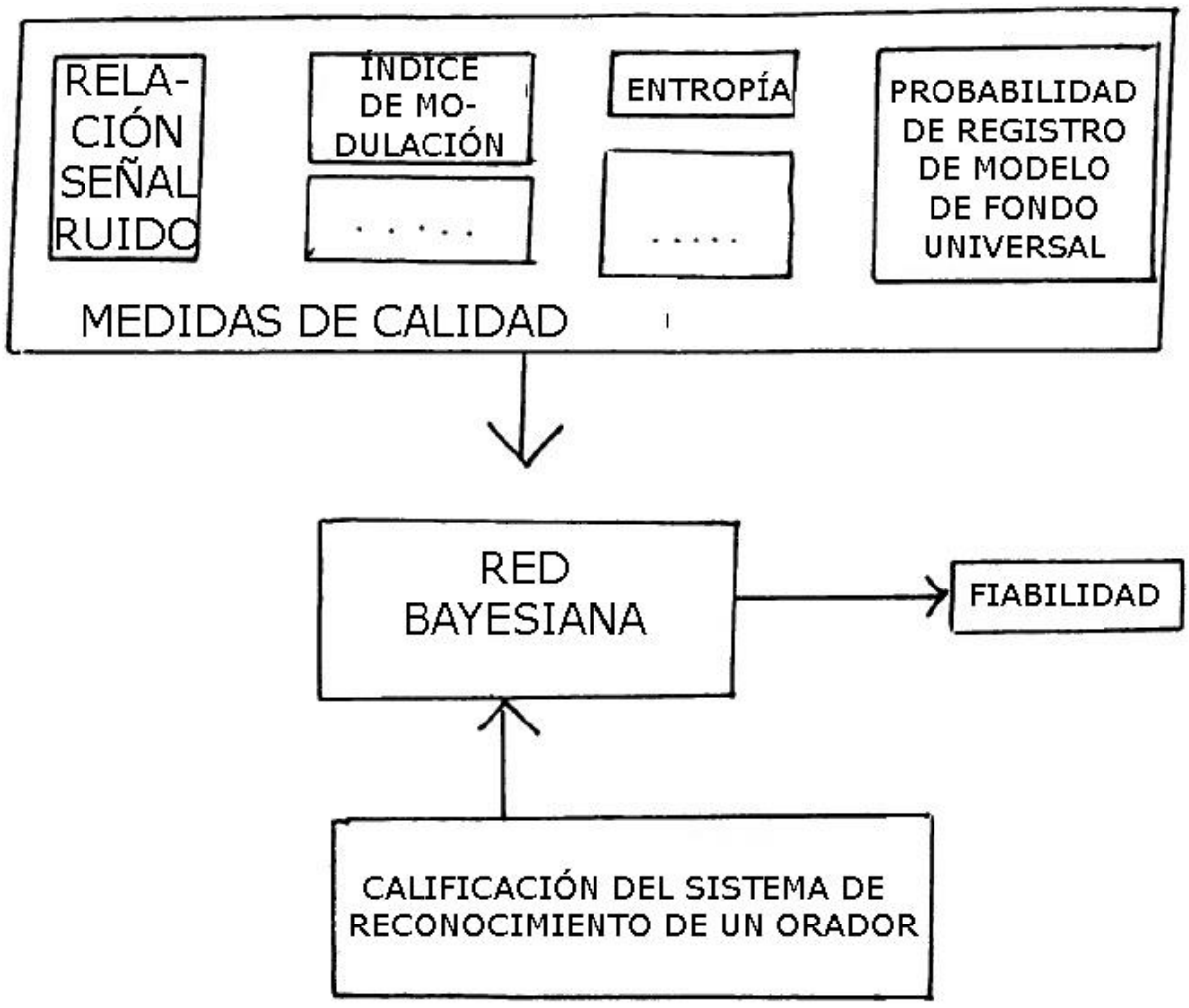
RELA-
CIÓN
SEÑAL
RUIDO
ÍNDICE
DE MO-
DULACIÓN
. . . . .
ENTROPÍA
. . . . .
PROBABILIDAD
DE REGISTRO
DE MODELO
DE FONDO
UNIVERSAL
MEDIDAS DE CALIDAD
RED
BAYESIANA
FIABILIDAD
CALIFICACIÓN DEL SISTEMA DE
RECONOCIMIENTO DE UN ORADOR

Fig. 3

Fig. 4a

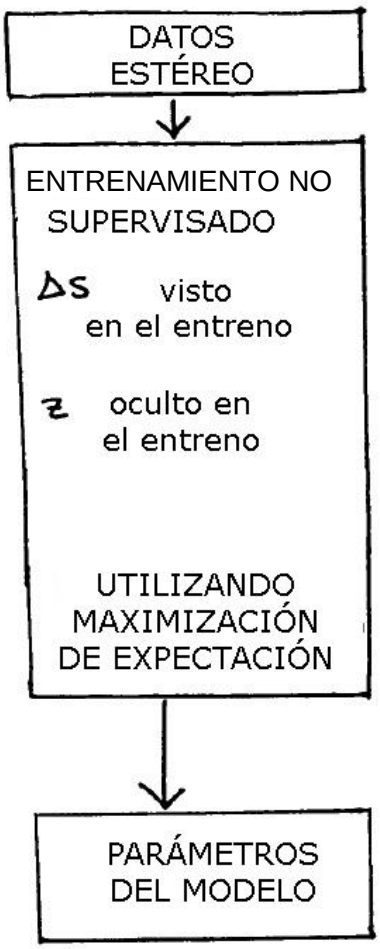

Fig. 4b

Fig. 4c

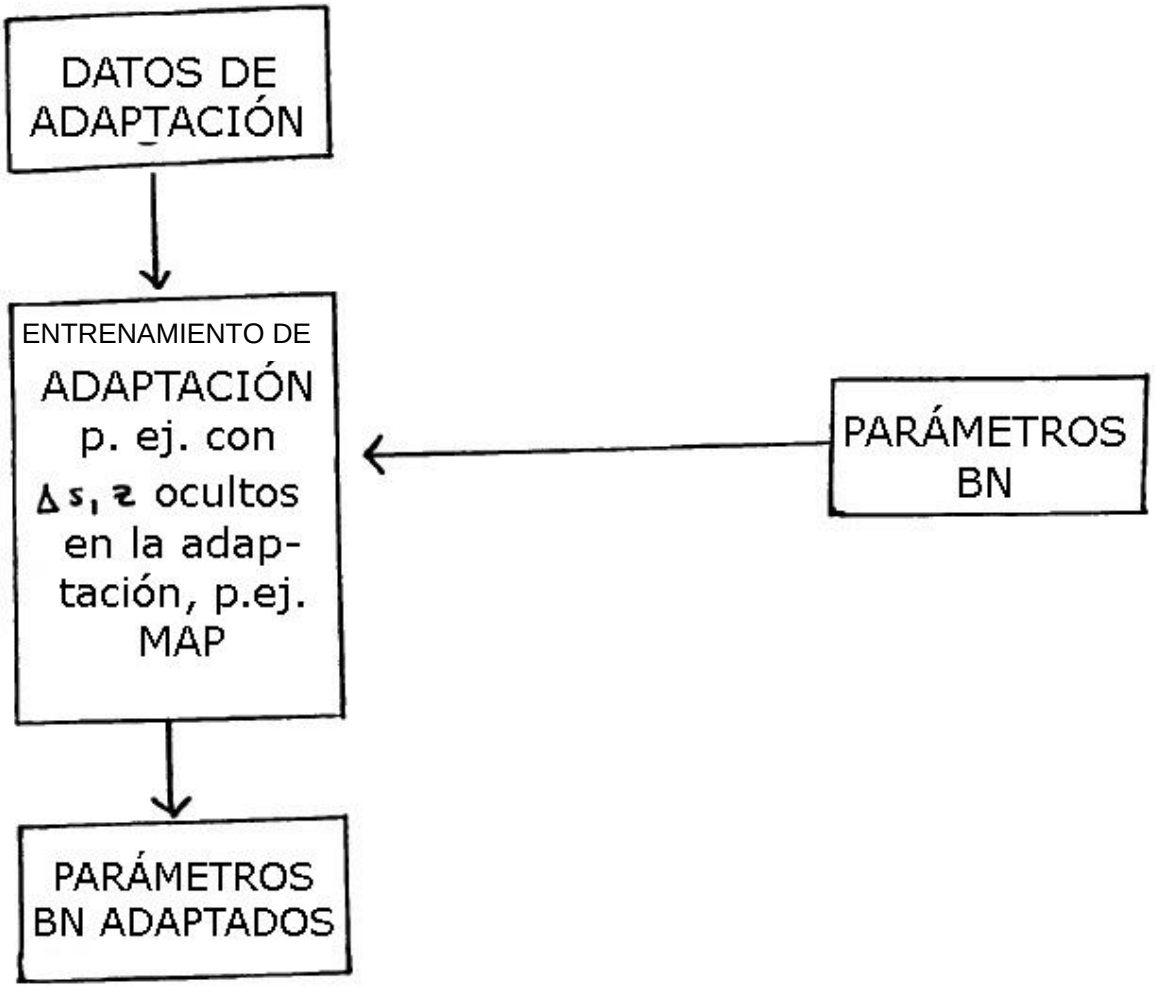
DATOS DE ADAPTACIÓN
ENTRENAMIENTO DE ADAPTACIÓN p. ej. con Δs, z ocultos en la adaptación, p.ej. MAP
PARÁMETROS BN
PARÁMETROS BN ADAPTADOS

Fig. 5

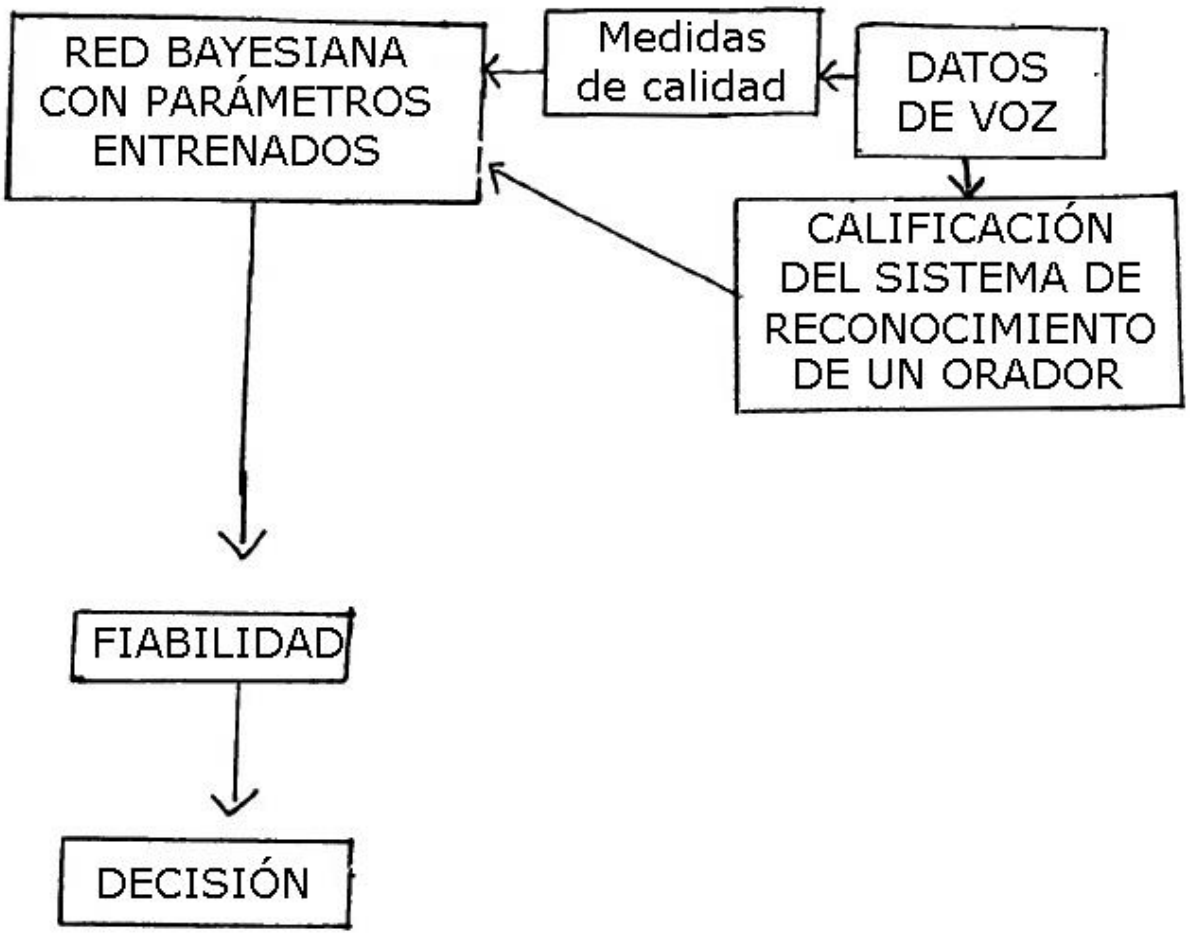
RED BAYESIANA CON PARÁMETROS ENTRENADOS
Medidas de calidad
DATOS DE VOZ
CALIFICACIÓN DEL SISTEMA DE RECONOCIMIENTO DE UN ORADOR
FIABILIDAD
DECISIÓN

Fig 6

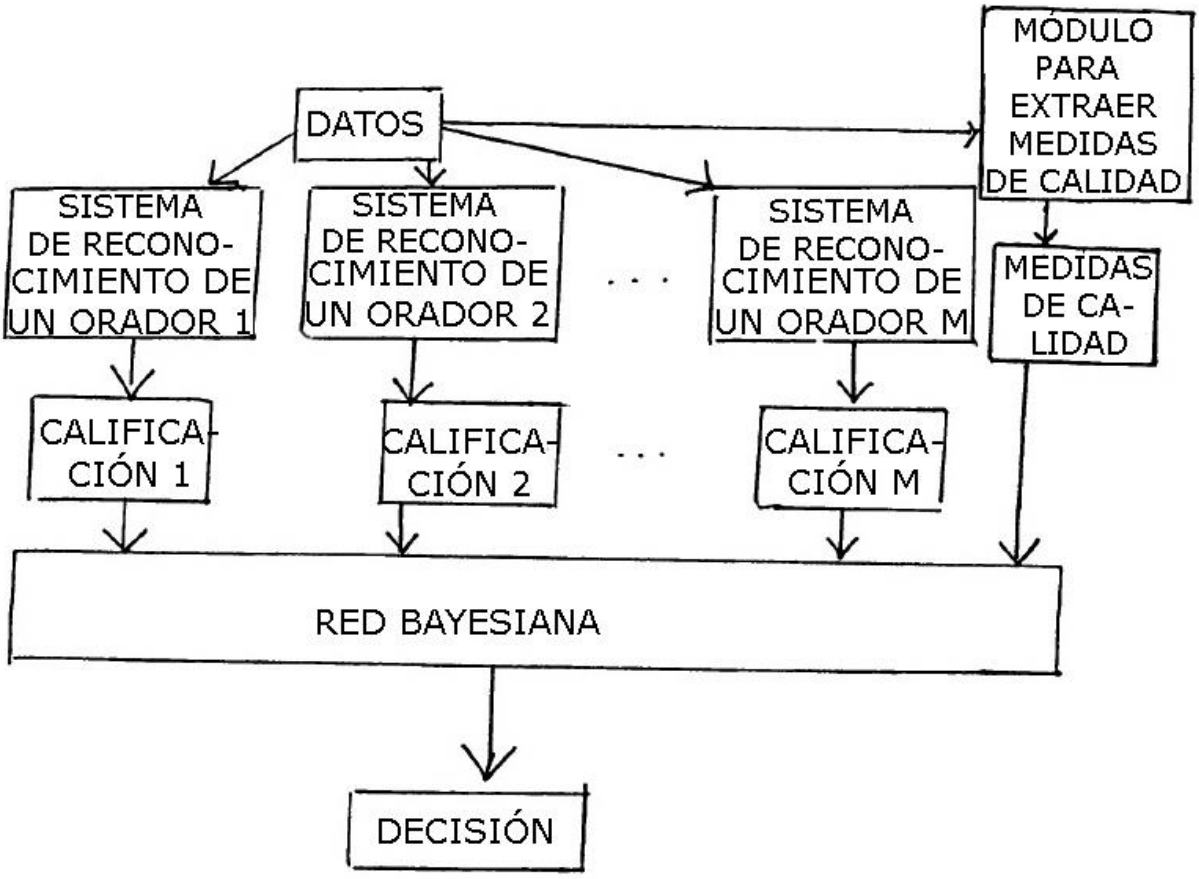
DATOS
MÓDULO PARA EXTRAER MEDIDAS DE CALIDAD
SISTEMA DE RECONO-CIMIENTO DE UN ORADOR 1
SISTEMA DE RECONO-CIMIENTO DE UN ORADOR 2
. . .
SISTEMA DE RECONO-CIMIENTO DE UN ORADOR M
MEDIDAS DE CA-LIDAD
CALIFICA-CIÓN 1
CALIFICA-CIÓN 2
. . .
CALIFICA-CIÓN M
RED BAYESIANA
DECISIÓN

Fig. 7

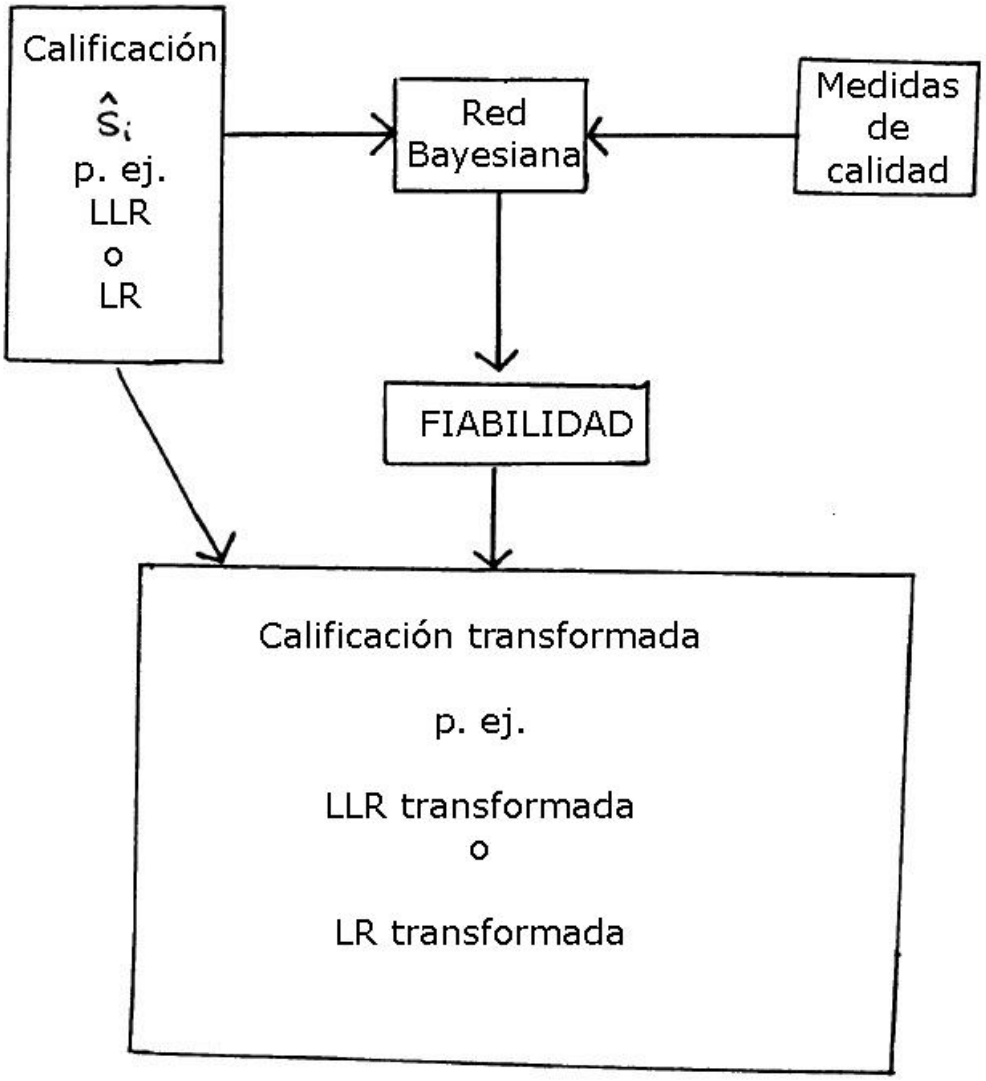
Calificación
$\hat{s}_i$
p. ej.
LLR
o
LR
Red Bayesiana
Medidas de calidad
FIABILIDAD
Calificación transformada
p. ej.
LLR transformada
o
LR transformada

Fig. 8