BATVOX: SISTEMA AUTOMÁTICO DE RECONOCIMIENTO DE LOCUTOR

BEATRIZ GONZÁLEZ SIGÜENZA
Agnitio
bgonzalez@agnitio.es

RESUMEN

El objetivo del presente documento es el de dar una visión general sobre el sistema de reconocimiento automático de locutor BATVOX y describir las líneas fundamentales de su funcionamiento. Para ello haremos un recorrido por todos los puntos de interés alrededor del funcionamiento del sistema. En primer lugar repasaremos, a grandes rasgos, cuales son los fundamentos teóricos de funcionamiento de la tecnología que usa BATVOX para realizar sus cálculos y, en segundo lugar, trataremos cuales son los procesos fundamentales que sufre un archivo de audio cuando es introducido en el sistema. Esto implica que veremos el proceso de extracción de características, parametrización y entrenamiento de los archivos de audio por parte del sistema.

Palabras clave: *Independiente de texto y lengua*.

ABSTRACT

The aim of this work is to provide a general vision on the automatic speaker recognition system BATVOX, and to describe the blueprint of its functioning. For this purpose, we will give an overview of the main features of the system. We will recall briefly the theoretical bases of the technology used by BATVOX to perform its calculations, and we will mention the basic processes that undergoes an audio file when it is introduced in the system. This implies that we will see how the system carries out the characteristics extraction process, the parameterization and the training of the audio files.

Keywords: *Text Independent and Language Independent.*

1. CONCEPTOS BÁSICOS

Vamos a ver brevemente una serie de conceptos necesarios para la comprensión de los procesos realizados por BATVOX. En particular, veremos las características básicas que posee la voz, veremos que componentes fundamentales la componen y cuales tienen importancia a la hora de realizar nuestro análisis de la voz.

2. NATURALEZA DE LA VOZ

Lo primero que podemos observar de la voz es cómo se comporta a lo largo del tiempo, lo que se denomina naturaleza temporal de la señal.

La señal de voz es una señal que a largo plazo (en el orden de segundos) debe ser considerada no estacionaria, puesto que sus características temporales se ven sometidas a constantes fluctuaciones, como se puede ver en la figura 1 en la que se representa una locución correspondiente a una frase de duración aproximada de 5 segundos.

Figura 1. *Locución de 5 segundos.*

Si en lugar de realizar un análisis a largo plazo (segundos), focalizamos nuestro estudio sobre tramos cuya duración esté un orden de magnitud por debajo (cientos de milisegundos) persistirá este carácter poco estacionario.

Sin embargo, llama la atención el hecho de que las transiciones entre sonidos son siempre progresivas, sin producirse de forma brusca.

La no estacionalidad de la señal hablada es una característica que se encuentra a largo (segundos) y a medio plazo (cientos de milisegundos). Sin embargo, en duraciones a «corto plazo», un orden de magnitud por debajo (decenas de milisegundos), la señal se comporta como quasi-estacionaria. Esto se puede ver, por ejemplo en la siguiente forma de onda de una vocal con una duración de 80 ms (figura 2).

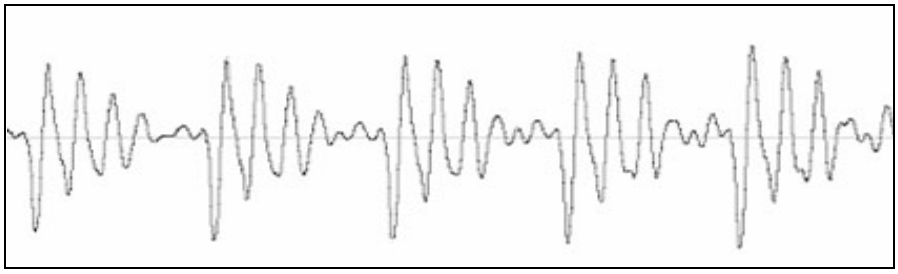

Figura 2. *Locución de una vocal de 80 ms.*

De la ilustración anterior se extrae además una conclusión importante, la señal de voz presenta, a corto plazo, una apariencia pseudo-periódica.

Esta característica, sin embargo no es generalizable, ya que es posible encontrar tramos a corto plazo con apariencia ruidosa (figura 3):

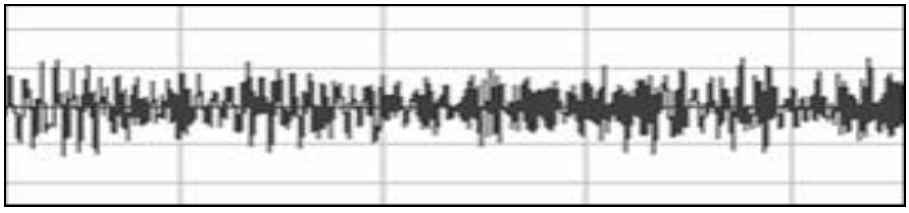

Figura 3. *Tramo sordo de apariencia ruidosa.*

A pesar de la constatación de la pseudo- estacionalidad de las señales habladas a corto plazo, llama la atención, como se acaba de mostrar, que existen tramos hablados de naturaleza distinta, puesto que mientras unos presentan un carácter periódico, otros, sin embargo, tienen una apariencia ruidosa.

Esta constatación permite realizar una clasificación genérica de los sonidos hablados en función de su naturaleza, como:

1. Sonoros, sonidos de carácter periódico

2. Sordos, sonidos de carácter ruidoso

3. CARACTERIZACIÓN ESTADÍSTICA DE LA VOZ

Desde el punto de vista estadístico es posible obtener algunas características generales de la voz entre las que destacan (González Rodríguez *et alii*, 2008):

1. *Distribución de los niveles de amplitud* que se puede ver en la figura 4. En su eje de abscisas tenemos el nivel de amplitud en dB, respecto al valor RMS. Este valor, denotado como 0 dB corresponderá a un valor absoluto de nivel de presión sonora (NPS) de unos 60 dB. En el eje de ordenadas tendremos la frecuencia (expresada en porcentaje) con la que el total de señal hablada analizada se encuentra por encima de un determinado nivel de amplitud (figura 4):

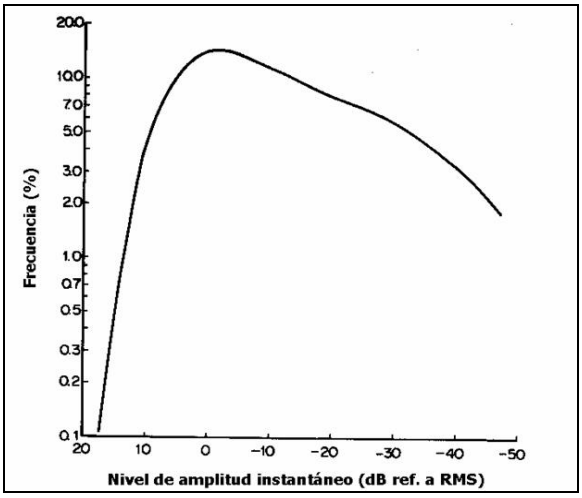

Figura 4. *Distribución de los niveles de amplitud.*

De la gráfica dela figura 4, resulta interesante extraer las siguientes conclusiones:

a. Los niveles máximos de amplitud se dan a +18dB con una frecuencia de aparición de tan solo el 0.1%.

b. Los niveles mínimos se dan a -50 dB, con frecuencia del 2% de la señal analizada

c. El margen dinámico extremo de la voz, relación de los niveles máximo y mínimo, es por tanto de unos 65 dB.

d. El margen dinámico de un hablante en situaciones de normalidad se sitúa entre 40 y 50 dB.

2. *Distribución de la frecuencia fundamental*: Un parámetro de importancia significativa es la distribución estadística de la frecuencia fundamental. Además, será conveniente caracterizar por separado a hombres y mujeres, puesto que sus cuerdas vocales tienen características fisiológicas diferenciadas. Las gráficas de la figura 5 muestran la distribución de frecuencia fundamental en hombres y mujeres. Como ya se ha mencionado previamente, la frecuencia fundamental es un parámetro que varía a lo largo del tiempo, produciendo la entonación característica de un determinado mensaje. De esta forma, las gráficas muestran la distribución de frecuencia instantánea, calculada de forma localizada

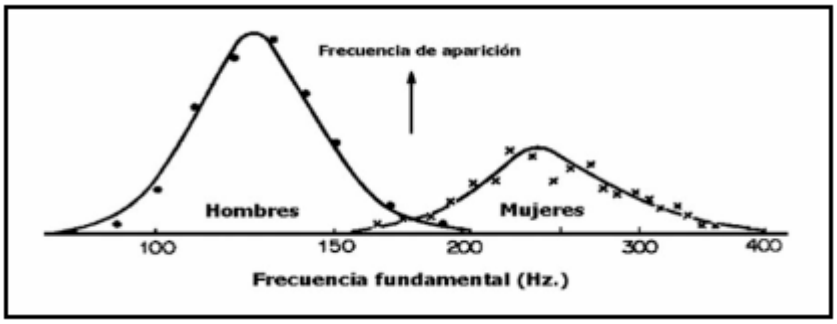

Figura 5. *Distribución de la frecuencia fundamental.*

A partir de la figura 5, podemos reseñar una serie de características destacadas:

a. Las distribuciones de frecuencia fundamental promedio siguen una función de densidad de probabilidad gaussiana, tanto en hombres como en mujeres.

b. La media de frecuencia fundamental en hombres se sitúa en torno a los 125 Hz, mientras que para las mujeres se sitúa en 250 Hz.

c. La dispersión de valores es menor en hombres que en mujeres.

d. Los valores extremos se sitúan entre los 80 y 200 Hz para los hombres y entre 150 y 400 para las mujeres.

4. PARAMETRIZACIÓN

En el desarrollo de aplicaciones de voz será necesaria la reducción de la cantidad de información disponible, así como la extracción de dicha información en dominios donde ésta sea suficientemente robusta e independiente.

En este sentido se propone la extracción de parámetros a partir de muestras de la señal, con el mencionado objetivo doble de reducir la cantidad de información a procesar y de expresar dicha información en dominios más adecuados. A esta extracción es a la que denominaremos *parametrización*.

Como hemos visto anteriormente, la señal de voz muestra características pseudo-estacionarias solo a corto plazo, en órdenes de decenas de milisegundos. Por consiguiente, si se desea aplicar técnicas de análisis y tratamiento de voz, debemos limitar el segmento a procesar en este orden de magnitud. Esto da origen al denominado *análisis localizado* de la señal, que obligará al uso de tramas de voz de la duración reseñada. El mecanismo que nos permite, dada una señal de voz, realizar un análisis localizado mediante el uso de tramas consecutivas se denomina *enventanado de la señal.*

5. SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR

Obviamente todos los procesos que hemos descrito obtienen como resultado final una serie de modelos que nos sirven para representar matemáticamente las

características de las diferentes voces. Pero el objetivo último de estos procesos no será el de la obtención de los modelos en sí mismos, sino que será la de poder realizar lo que se denomina *Tareas de reconocimiento*.

En ellas, el objetivo será el de determinar si uno o varios audios cuya procedencia nos es desconocida son los más parecidos y/o pertenecen a uno o varios modelos de locutores que poseeremos y que habrán sido obtenidos con anterioridad de locutores identificados.

Todos estos procesos se realizarán en sistemas denominados *Sistemas De Reconocimiento Automático de Locutor*.

6. PRINCIPIO Y MODOS DE FUNCIONAMIENTO

Para realizar las tareas encomendadas, este tipo de sistemas suelen trabajar en tres modos o fases distintas.

1. *Modo de entrenamiento*: En esta fase, que puede ser realizada con el sistema en funcionamiento (on-line) o como un proceso independiente y anterior a la puesta en marcha del mismo (off-line), se obtiene la información necesaria (en forma de patrones o modelos) que se usará como valor de referencia correspondiente a cada uno de los usuarios del sistema.

2. *Modo de prueba, funcionamiento o servicio*: Esta será la fase de utilización del sistema; en ella, a partir de nuevas señales habladas, el sistema tomará decisiones acerca de la identidad del hablante.

3. *Modo de actualización*: Durante la vida útil del sistema, éste deberá ser capaz de incorporar nuevos locutores, dar de baja a usuarios y, opcionalmente, actualizar o mejorar los modelos y referencias correspondientes a los usuarios presentes en el sistema.

A continuación se presenta en la figura 6 el diagrama de bloques de un sistema de reconocimiento genérico de locutores:

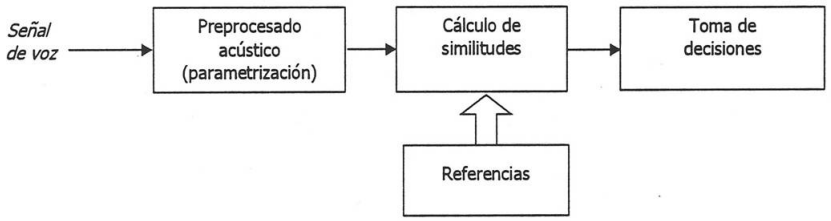

Figura 6. *Diagrama de bloques de un sistema de reconocimiento genérico de locutores.*

a. El sistema parte de una locución procedente de un locutor no clasificado.

b. El módulo llamado *Preprocesado Acústico*, convertirá la señal acústica de entrada en una serie de vectores de características que extraigan de forma eficiente la información de locutor presente en la señal de voz. Opcionalmente se podrán incluir funciones para dotar de mayor robustez acústica el sistema.

c. *El módulo de patrones/referencias* dispondrá de patrones o referencias correspondientes a los distintos locutores conocidos por el sistema (usuarios) y obtenidos en la fase de entrenamiento.

d. *El módulo de cálculo de similitudes*, una vez obtenidos los vectores de características correspondientes a la señal de voz de entrada, y teniendo disponibles los modelos o patrones correspondientes a los distintos locutores, calculará el parecido o similitud entre la realización acústica de entrada y cualquiera de los modelos conocidos por el reconocedor.

e. El módulo final de *toma de decisiones*, a partir de los valores de similitud obtenidos, deberá tomar una decisión acerca de la identidad del locutor que ha generado la locución de entrada.

Si bien este es el esquema que seguirán todos los sistemas de reconocimiento en general, se podrán diferenciar entre sí dependiendo de diversas características propias de cada sistema. Entre ellas la más destacada es su comportamiento ante el texto pronunciado y su dependencia ente él.

Habrá sistemas que sean completamente *dependientes de texto* que consistirán en que tanto el entrenamiento como la prueba son completamente idénticas (típicamente la pronunciación de una clave individual o una común para el sistema). En ese caso el sistema solo tendrá que realizar una comparación entre realizaciones diferentes del mismo «texto».

Sin embargo, habrá otras aplicaciones que permitirán *independencia del texto* pronunciado en mayor o menos medida. Es decir, no coincidirá la locución a pronunciar en la fase de entrenamiento con la pronunciada en el reconocimiento. Dependiendo del grado de independencia obtenido, tendremos sistemas que poseen un vocabulario fijo y finito, otros basados en sucesos (son capaces de detectar la pronunciación de una palabra concreta en un marco no limitado) y por último los que son independientes sin ningún tipo de restricción. El sistema BATVOX y su tecnología de reconocimiento se encuentran entre estos últimos.

7. TAREAS DE RECONOCIMIENTO

En este apartado vamos a describir brevemente en qué consisten las diferentes Tareas de Reconocimiento que puede realizar un Sistema Automático de Reconocimiento de Locutor. Dependiendo cual sea su objetivo podremos clasificar las tareas en dos grandes grupos que describiremos a continuación. Se trata de las tareas de identificación y las tareas de verificación.

Como veremos también, a su vez, estas tareas se pueden subdividir dependiendo de la forma y métodos que usemos a la hora de realizarlas

8. IDENTIFICACIÓN

El objetivo de una identificación de locutores es el de clasificar una señal de voz, cuyo origen no conocemos, como perteneciente a uno de entre un conjunto de N posibles locutores.

Dentro de estos sistemas, debemos diferencias dos posibles casos:

1. *Identificación en conjunto cerrado*: en este caso, el resultado del proceso es una asignación de identidad a uno de los locutores modelados por el

sistema, y conocidos como «usuarios». Existen por tanto, N posibles salidas posibles.

2. *Identificación en conjunto abierto*: Aquí debemos considerar una posibilidad adicional a las N del caso anterior, y es que el locutor que pretende ser identificado no pertenezca al grupo de usuarios, con lo que el sistema de identificación debería contemplar la posibilidad de no clasificar la locución de entrada como perteneciente a las N posibles.

En principio, la comparación directa en un sistema automático de una serie de modelos frente a un audio cualquiera solo será fiable cuando podamos garantizar que las características de todos los modelos son similares (más adelante en este manual veremos que variables hay que tener en cuenta para comprobar la correspondencia de esas características).

En caso contrario, cabe el riesgo de que la identificación realizada pueda ser errónea. En un sistema automático hay una serie de variables, como pueda ser, por ejemplo, el canal de grabación (más adelante se verán todas con detalle) que provocan que en el caso de que haya modelos de diferentes características (por ejemplo, uno grabado en canal microfónico y otro en canal telefónico) los resultados puedan adulterarse (en el ejemplo anterior, un test grabado en canal telefónico tenderá a puntuar más alto con los modelos telefónicos aunque no sean de la misma persona).

Por esa razón, si no podemos asegurar el extremo de que todos los modelos son iguales lo más recomendable es realizar una «normalización». Este procesose encargará de compensar las puntuaciones para permitir compararlas.

De esta manera tendremos que también se podrá clasificar las identificaciones entre *simples* y *normalizadas*

9. VERIFICACIÓN

En la verificación de locutores, en contraposición a la identificación se reciben dos entradas. Una de ellas es la señal de voz a verificar y la otra es una solicitud de identidad, que puede ser realizada de diversas formas. De este modo, las dos únicas

salidas o decisiones del sistema son la aceptación o rechazo de la hipótesis de que ambas locuciones pertenezcan a la misma persona. La decisión de aceptar o rechazar la locución de entrada como correspondiente al locutor solicitado dependerá de si el valor del parecido o probabilidad obtenido supera o no un determinado umbral de decisión.

Así será preciso establecer en el proceso de entrenamiento, junto a los patrones o modelos de cada uno de los locutores usuarios del sistema, el conjunto de valores o umbrales, uno por locutor, que permitirán al sistema tomar sus decisiones.

Existirán diversas tácticas y técnicas para abordar el problema del establecimiento de umbrales. Una de ellas (que es además la utilizada por la tecnología de reconocimiento de BATVOX) es la denominada de Relaciones de Verosimilitud que será descrita en el apartado siguiente

10. LIKELIHOOD RATIOS (LR)

Las Relaciones de Verosimilitud o LRs (*Likelihood Ratios*) (González-Rodríguez *et alii*, 2006) son una aproximación a la tarea de verificación introduciendo para ello las teorías bayesianas de apoyo a la decisión.

Esta aproximación bayesiana esta fuertemente establecida en otras disciplinas forenses como el análisis del ADN y está orientada a su utilización en un entorno judicial.

Su enunciación parte de la clara separación de los roles de juez/jurado que se encargan de las tareas de decisión y el de el científico que se encarga del análisis de los datos e interpretación de los resultados para proporcionar elementos que ayuden a la decisión.

Por ello, el resultado de una verificación realizada a través de LRs nunca será una decisión categórica, sino una probabilidad. O más concretamente, un cociente de probabilidades.

Estas probabilidades serán las que relacionan afirmativa o negativamente las identidades de la voz indubitada y la correspondiente dubitada y darán un refuerzo o debilitamiento de la hipótesis inicial.

Los conceptos que manejaremos en este entorno son:

1. La hipótesis inicial, es decir, que la voz dubitada pertenece al sospechoso. La llamaremos C.

2. La observación del experto forense (en este caso, el sistema de reconocimiento automático de locutor) que denominaremos E.

3. Las circunstancias del caso, que son todo aquello ajeno a la observación del experto, que, sin embargo influye en él. La llamaremos I.

Con estos elementos poseeremos lo que se denomina *probabilidad a priori* y que expresaremos como aparece en el apratado a de la figura 7. Es decir, será las probabilidades de que la voz dubitada pertenezca al sospechoso teniendo en cuenta todos los elementos y circunstancias del caso a excepción de la observación del experto forense.

Obviamente, estas probabilidades son muy importantes e influyentes a la hora de tomar una decisión. De hecho, estas probabilidades pueden eliminar o corroborar completamente la hipótesis inicial por si misma sin necesidad de ningún estudio de las voces por parte de un experto.

También, como es lógico, existirá lo que denominaremos *probabilidades a posteriori.* Estas serán las probabilidades de que la hipótesis inicial sea correcta, pero en este caso, teniendo en cuenta, no solo las circunstancias del caso, sino también la observación de las voces por un experto forense. De esta manera, expresaremos esa probabilidad como aparece en b de la figura 7.

Como vemos, el paso de una probabilidad a otra lo único que incluye es el añadido del análisis de un experto forense. Esto en la práctica, se reduce a la ponderación de la probabilidad a priori por una constante. Esta constante representará el resultado de la observación del experto forense y como resultado nos dará la probabilidad a posteriori. A esta constante de ponderación es a la que denominamos Relación de Verosimilitud o LR. Véase el apartado c en la figura 7.

Ese LR, como hemos dicho, representará toda la información computada por el experto en su observación y se haya a través del cociente entre dos probabilidades.

La primera de ellas será la probabilidad de que, dada la observación, la hipótesis inicial sea la correcta y la otra será justo la contraria, es decir, que dada la observación la hipótesis inicial sea incierta (o lo que es lo mismo, que sea correcta la hipótesis contraria, véase d en la figura 7). Esto se expresa como se observa en e de la figura 7:

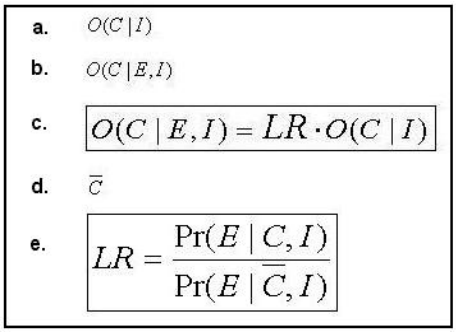

Figura 7. *Fórmulas.*

De esta manera, es obvio que si lo más probable es que la voz dubitada pertenezca al sospechoso el LR tendrá un valor mayor que uno. De igual manera, si por el contrario, es más probable que la voz dubitada no pertenezca al sospechoso el LR será menor que uno. Obviamente, cuanto más probable sea una hipótesis frente a la otra el LR se hará mayor o menor en cada caso y su rango irá desde cero al infinito.

De esta manera la función del LR no será más que reforzar o debilitar las probabilidades que existían a priori de que fuera cierta la hipótesis inicial, pero nunca sustituirla.

11. CONCLUSIÓN

BATVOX es la mejor herramienta de reconocimiento automático de locutor para laboratorios de acústica que existe en el mercado. Ha sido presentado como prueba a juicio en varios países en Europa, Asia y América latina. Diferentes científicos corroboran la utilidad de este sistema y garantizan nuestro trabajo.

12. REFERENCIAS BIBLIOGRÁFICAS

GONZÁLEZ-RODRIGUEZ, J.; D. TORRE TOLEDANO y J. ORTEGA-GARCÍA (2008): «Voice biometrics», en A. K. Jain., P. Flynn y A. A. Ross (eds): *Handbook of Biometrics*, capítulo 8. pp. 151-170.

GONZALEZ-RODRIGUEZ, J.; A. DRYGAJLO; D. RAMOS-CASTRO; M. GARCIA-GOMAR y J. ORTEGA-GARCIA (2006): «Robust Estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition», *Computer Speech and Language*, vol. 20, pp. 331-335.