Computación Blanda

Soft Computing

**Autor:** Nicolás Aguirre Espinosa

IS&C, Universidad Tecnológica de Pereira, Pereira, Colombia

**Correo-e:** [nicolas.aguirre@utp.edu.co](mailto:nicolas.aguirre@utp.edu.co)

**Resumen**

El Reconocimiento Automático de Locutor o Reconocimiento de Locutores (ASR), es un campo de investigación de creciente relevancia que día a día se gana más adeptos. El desarrollo de mejores algoritmos y de modelados más precisos, junto con la aparición de sistemas informáticos más potentes y asequibles, posibilita la integración de los sistemas de dialogo hombre-máquina a través de la voz en numerosos ámbitos de la sociedad actual.

El objetivo general de este documento es presentar una visión general de los principales aspectos y punto de interés relacionados con el Reconocimiento del Automático del locutor, así como sus principales aplicaciones y dificultades en el ámbito práctico y teórico. A su vez se describirán los principales aspectos a tener en cuenta a la hora de implementar el algoritmo de K-NN o de N vecinos cercanos.

**Palabras claves:** inteligencia artificial, reconocimiento de locutores, reconocimiento de voz, informática, reconocimiento automático del locutor, ASR, algoritmo K-NN.

**Abstract**

Automatic Announcer Recognition or Announcer Recognition (ASR) is a field of research of increasing relevance that is gaining more followers every day. The development of better algorithms and more precise modeling, together with the emergence of more powerful and affordable computer systems, makes it possible to integrate human-machine dialogue systems through voice in many areas of today's society.

The general objective of this document is to present an overview of the main aspects and points of interest related to the Automatic Speaker Recognition, as well as its main applications and difficulties in the practical and theoretical field. In turn, the main aspects to be taken into account when implementing the K-NN or N close neighbors algorithm will be described.

**Keywords:** artificial intelligence, announcer recognition, voice recognition, computing, automatic speaker recognition, ASR, K-NN algorithm.

**Introducción:**

Uno de los aspectos que con el pasar del tiempo ha recobrado mayor importancia es el desarrollo de equipos, programas y sistemas de computación que contemplen y mejoren la comunicación con el usuario.

Hace ya tiempo que se vienen desarrollando interfaces hombre- máquina controlados por voz que permitan sustituir en ciertas aplicaciones a las interfaces tradicionales basadas en teclados, paneles y dispositivos similares. Los primeros avances en este campo fueron a inicios de los años 50 con Davis, Bidulph y Balashek, de los laboratorios Bell fabricando el primer reconocedor capaz de discriminar con cierta precisión los diez dígitos ingleses pronunciados de forma aislada por un único lector. El dispositivo era totalmente electrónico. Los primeros trabajos que hacen uso de tecnología informática comienzan a aparecer en a finales de los años 50 e inicios de los años 60. Deves y Mathews introducen el concepto de normalización temporal no lineal, que permite la comparación de parámetros de palabras iguales pronunciadas a distinta velocidad.

A partir de estas fechas comienza el auge por la investigación de este campo, dando como resultado un sin fin de trabajos, enfocados principalmente en el reconocimiento de palabras aisladas; en poco tiempo, se avanzó notablemente en sistemas capaces de reconocer de forma precisa frases cualesquiera, pronunciadas por un lector cualquiera, de forma continua. Motivados con los resultados obtenidos alrededor de este campo, se lanzan grandes proyectos de investigación en todo el mundo en los que se pretendía llegar a las menores restricciones gramaticales posibles de las frases a reconocer, así como del léxico utilizado

Aunque muchos de los ambiciosos proyectos lanzados por la época no llegaron realmente a alcanzar los objetivos que se habían planteado, este proceso trajo consigo múltiples aportes derivados, entre ellos, el desarrollo de un mejor conocimiento alrededor de los mecanismos del habla y de las limitaciones de los sistemas automáticos de reconocimiento.

Los avances que se producen en el ámbito de las tecnologías del habla son día a día más significativos, pero en este documento nos enfocaremos en el Reconocimiento Automático de Locutor, rama de la inteligencia artificial que consiste en identificar automáticamente a una persona a través de su voz. El hecho de poder distinguir un locutor de otro está relacionado mayoritariamente con las características fisiológicas y los hábitos lingüísticos de cada uno ellos. El reconocimiento conlleva un procesado de audio que permite extraer este conjunto de rasgos inherentes al locutor y la posterior búsqueda de posibles coincidencias mediante un proceso de reconocimiento de patrones.

Los sistemas de reconocimiento automático de locutor se pueden clasificar en dos tipos principales:

•Identificación de locutores

•Verificación o detección de locutores

**Identificación de locutores**

En la identificación de locutor, el locutor no aporta información sobre su identidad y es el sistema el que determina quién es a partir de su voz dentro de un conjunto de posibles candidatos o, si se trata de identificación en conjunto abierto, si el locutor es conocido o no por el sistema. En un sistema de identificación el sistema suele recibir una o varias muestras de voz y las contrasta con una base de datos con voces cuyas identidades son conocidas. Luego, el sistema asigna una puntuación de semejanza a cada una de estas identidades, obteniendo puntajes más altos los de aquellas personas cuyas voces tienen mayor coincidencia con la muestra con la que se están comparando.

**Verificación o detección de locutores**

La tarea de los sistemas de verificación de locutor es determinar si el locutor es o no quién dice ser. La decisión es binaria; el sistema recibe una grabación con la voz del locutor y la identidad proclamada por este y luego el sistema da como salida el éxito o fracaso de esta verificación.

En muchos casos se puede considerar un tercer tipo que sería el seguimiento y agrupamiento de locutores que consiste en etiquetar qué locutor está hablando en un segmento de voz y cuándo se producen cambios de locutores.

En algunos campos como lo es el forense es común llevar a cabo primeramente un proceso de identificación para crear una lista de identidades con alta probabilidad de coincidencia. Luego, un proceso de verificación permite llegar a un resultado final, con una única identidad definida.

Según el contenido de la señal de voz empleada las modalidades de reconocimiento se clasifican en independientes del texto y dependientes del texto.

**Independientes del texto**

Éste es el caso general, donde el sistema no conoce el texto hablado por la persona. Este modo de operación es obligatorio en aquellas aplicaciones donde el usuario no sabe que es evaluado por un sistema de reconocimiento, como sucede en algunas aplicaciones forenses, o al simplificar el uso de un servicio donde la identidad es inferida para establecer un diálogo hombre/maquina, como se hace en ciertos servicios bancarios. Esto permite mayor flexibilidad, pero también aumenta la complejidad del problema. Si es necesario, para el reconocimiento del discurso se puede proporcionar el conocimiento del texto hablado. En este modo uno puede utilizar indirectamente la coocurrencia típica de las palabras del locutor, para evaluarlo desde una probabilidad gramatical. Este modelo de coocurrencia se conoce como n- gramas, y permiten determinar la probabilidad de que n palabras consecutivas sean pronunciadas por un locutor.

**Dependientes del texto:**

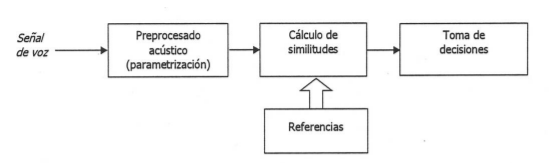
Este modo de operación implica que el sistema conoce el texto hablado por las personas. Puede ser un texto predefinido o inducido. En general, El conocimiento del texto hablado nos permite mejorar el funcionamiento del sistema con respecto a la categoría anterior. Este modo es de gran importancia en las aplicaciones donde se hacen estrictos controles de entrada al usuario de un sistema, o en las aplicaciones donde una unidad de diálogo va orientando al usuario.

En un sistema de verificación de locutor (y en general, en cualquier sistema de reconocimiento de voz), siempre nos vamos a encontrar con tres fases o etapas:

**Fase 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.

**Fase de Prueba o Test:** 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.

**Fase de actualización o mantenimiento:** 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.



*Diagrama de sistema genérico de reconocimiento de locutores*

Como se presenta en la anterior figura el diagrama de bloques de un sistema de reconocimiento genérico de locutores consiste principalmente en los siguientes pasos:

1. El sistema parte de una locución procedente de un locutor no clasificado.
2. 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.
3. 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.
4. 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.
5. 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.

**Problemas de aplicación:**

A pesar de lo básico que aparente ser el problema de identificación de un locutor para los humanos, el estudio de la misma muestra de forma inmediata, una enorme complejidad. El principal problema que presenta el ASR es su carácter interdisciplinar, es decir, deber ser abordado a través de diferente técnicas y procedimientos en distintas áreas del conocimiento, no obstante, está no es la única dificulta en la implementación del ASR, también existen otros aspectos prácticos que se pueden agrupar en las siguientes categorías:

**Continuidad**: En el lenguaje natural no existen separadores entre las unidades, ya que no existen silencios, en algunos casos, ni entre las palabras.

**Dependencia del contexto:** Cada sonido elemental en los que se puede dividir el habla es modificado por el contexto en el que se encuentra.

**Variabilidad:** Se pueden distinguir dos tipos de variabilidad. La variabilidad intra- hablante está relacionada con las modificaciones introducidas por un mismo hablante sobre diferentes pronunciaciones de los mismos fonemas o palabras. La variabilidad inter-hablante se debe a aspectos relacionados con el locutor y el entorno, ya que la señal obtenida dependerá de los dispositivos utilizados en su captación, del entorno donde se obtiene y, principalmente, de aspectos anatómicos particulares del aparato fonador de cada hablante.

**Necesidades de almacenamiento:** Debido a las causas anteriores, se hace necesario procesar y almacenar grandes cantidades de datos.

**Estructuración:** La misma señal contiene información sobre varios niveles de descripción. De esta forma, una frase puede ser descrita a nivel semántico, sintáctico o fonético. Por otra parte, una señal de voz contiene información sobre el locutor que la emite. Así, es posible distinguir el sexo y la identidad de la persona a partir de la propia señal.

**Inexistencia de reglas de descripción y redundancia:** No existen reglas precisas capaces de describir los diferentes niveles en los que se presenta la información.

**Aplicaciones**

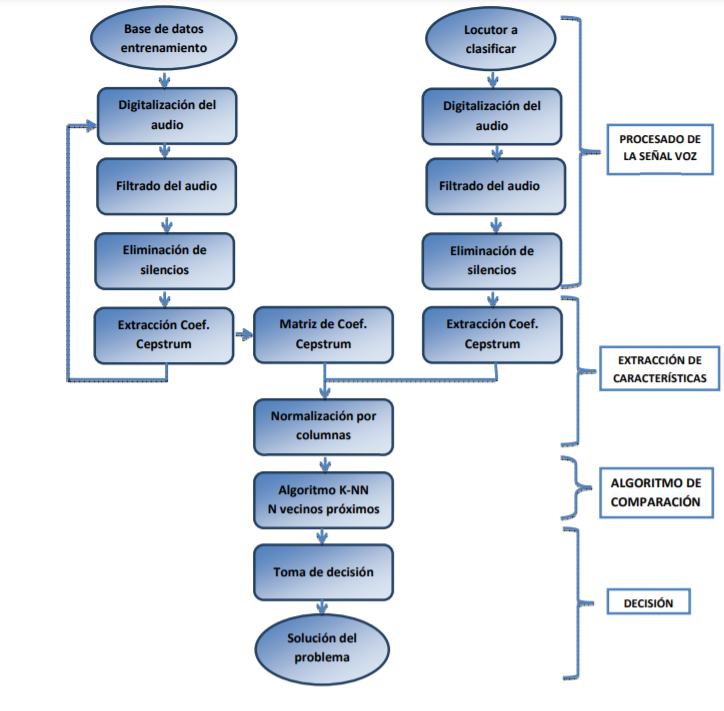
El desarrollo de tecnologías encargadas de reconocer automáticamente a una persona mediante su voz ha experimentado un creciente interés en los últimos años debido a sus múltiples aplicaciones, entre ellas están:

* **Control de acceso:** Acceso a instalaciones físicas, acceso a un ordenador.
* **Transacciones de autenticación:** Comercio electrónico, transacciones bancarias.
* **Servicio personalizado:** Aplicaciones de domótica.
* **Gestión de audio:** Indexación automática de contenidos de audio
* **Forense:** Identificación de personas a través de grabaciones para validar pruebas.

Como se hizo alusión anteriormente el ARS tiene una amplia variedad de algoritmos y formas de aplicación, en este caso nos centraremos en uno el algoritmo de aprendizaje supervisado K-NN o de N vecinos cercanos, consiste en asociar a cada trama de la matriz test su hablante más parecido basándose en los N vectores más similares de la matriz train. Para medir esa semejanza entre vectores de coeficientes se utiliza la distancia euclídea.

Tendremos por lo tanto un conjunto de valores que señalan con qué hablante se estima que se corresponde cada una de las tramas de sonido. Por último, sólo queda determinar a qué hablante se corresponde todo el audio, para lo cual bastará con elegir el locutor que más veces se haya estimado en el paso anterior.

El esquema de la figura a continuación muestra el algoritmo seguido durante todo el programa hasta determinar a qué hablante pertenece cada audio de test y en los siguientes apartados se hablará sobre cada uno de los bloques.



**1.Procesado de la señal de voz**El primer bloque del sistema consiste en la selección de los audios y un procesado de la voz que entra al sistema, con el objetivo de extraer sólo la información acústica relevante de los audios. Las señales de entrada pueden estar contaminadas con un ruido de fondo o con largos períodos sin habla que convenientemente deben ser eliminados. Esta función está dividida en tres pasos fundamentales: Selección de audios, reducción de ruido y supresión de silencios.

**2.Extracción de características**  La extracción de parámetros o características obtiene vectores de un determinado objeto que modelan su patrón. La voz puede simularse como un filtro vocal del cual puede representarse un espectro de las componentes frecuenciales de la voz y extraer información representativa.

**3.Algoritmo de comparación**

Para asignar una clase a un conjunto de parámetros se utiliza la clasificación de patrones. La asignación de clases se realiza para llevar a cabo una diferenciación entre subconjuntos de características.

Para los sistemas de ARS es necesario que los patrones que describen objetos de una misma clase presenten características similares. Hay distintos tipos de patrones:

* **Patrones vectoriales:** Codifican variables concretas significativas.
* **Patrones estructurados**: Codifican relaciones entre componentes del objeto o descriptores. Hay muchos tipos como por ejemplo árboles o cadenas.

**4.Distancia Euclídea**

La medida de distorsión más utilizada es la distancia Euclídea. Esta distancia se utiliza para el reconocimiento del hablante como método para calcular las diferencias existentes entre características. El resultado final de dicha comparación son valores numéricos que representan la distancia entre vectores de iguales dimensiones.

**5.Algoritmo K-NN:**

La agrupación de objetos atendiendo a sus características ha sido ampliamente estudiada debido a sus numerosas aplicaciones como aprendizaje de máquina.

El objetivo es reorganizar un grupo de objetos, en este caso pequeñas tramas de audio, los cuales tienen asociados vectores multidimensionales en grupos homogéneos, tales que los patrones de cada grupo son similares.

**Toma de decisión**

El algoritmo de comparación hace prácticamente todo el trabajo con respecto a la determinación de a qué clase pertenece cada objeto.

El algoritmo K-NN lo que cataloga son fracciones de audios y no el audio al completo, por lo que el paso final es llegar a una conclusión para catalogar dicho audio. En este caso la detección consistiría en elegir hablante final aquel que se ha correspondido más veces con las fracciones de sonido estudiadas.

**Bibliografía:**

*[1]* 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.*

[2] Beatriz González Sigüenza. *«BATVOX: Sistema automático de reconocimiento de locutor». PDF*

[*https://www.ub.edu/journalofexperimentalphonetics/pdf-articles/XVII-17.pdf*](https://www.ub.edu/journalofexperimentalphonetics/pdf-articles/XVII-17.pdf)

[3] Maestre R. José y Serje D. Christian (2007): *«Técnicas para reconocimiento automático de locutores». PDF*

[*https://biblioteca.utb.edu.co/notas/tesis/0040351.pdf*](https://biblioteca.utb.edu.co/notas/tesis/0040351.pdf)

*[4] «Sistemas de Verificación automática de locutor. PDF*

[*http://bibing.us.es/proyectos/abreproy/12054/fichero/MEMORIA%252F7.Cap%C3%ADtulo+2.pdf*](http://bibing.us.es/proyectos/abreproy/12054/fichero/MEMORIA%252F7.Cap%C3%ADtulo+2.pdf)

*[5] Eugenio Arévalo González (2011):*

*«Reconocimiento Automático de Locutor en Entornos Forenses basado en Técnicas de Factor Analysis aplicadas a Nivel Acústico». PDF*

*<http://arantxa.ii.uam.es/~jms/pfcsteleco/lecturas/20110714EugenioArevalo.pdf>*

*[6] Cristina Esteve Elizalde (2007):*

*«Reconocimiento de locutor dependiente de texto mediante adaptación de modelos ocultos de markov fonéticos». PDF*

[*https://pdfs.semanticscholar.org/780f/3a2d07df0eac52f56c63e8795f3a020fedd7.pdf*](https://pdfs.semanticscholar.org/780f/3a2d07df0eac52f56c63e8795f3a020fedd7.pdf)

*[7] Wikipedia Org.*

*«Reconocimiento de locutores».*

[*https://es.wikipedia.org/wiki/Reconocimiento\_de\_locutores*](https://es.wikipedia.org/wiki/Reconocimiento_de_locutores)