RECONOCIMIENTO DE ARCHIVOS DE AUDIO

RECOGNITION OF AUDIO FILES

Autor 1: Jose Daniel Velasquez
Universidad tecnológica de Pereira-Risaralda, Colombia
Correo-e: j.velasquez@utp.edu.co

***Resumen*— En este articulo veremos a que nos referimos cuando hablamos de reconocimiento de locutor o hablante, como se diferencia del reconocimiento de voz para texto. Los dos tipos de finalidades del reconocimiento del locutor, que son identificación y verificación, siendo esta ultima la mas parecida a la tecnología biométrica por voz. además de estos dos tipos de visiones de un sistema como este, también se dividen en dos dependiendo de si son dependientes o independientes de texto, siendo los dependientes los sistemas mas fáciles para extraer parámetros a la hora de analizar el espectro de voz. Los tipos de información que se pueden extraer de un audio dado, las implicaciones de esta y su dificultad de extracción. Los diferentes tipos de análisis de voz, los mas destacados aquellos que se basan en el espectro de este, y haciendo uso de la transformada de Fourier (En este caso específicamente FFT) como podemos sacar parámetros y las bases de estos métodos. Por ultimo el modo de selección del locutor mas apropiado partiendo de un banco de entrenamiento y un audio analizado, basado en probabilidades t el teorema de Bayes.**

Palabras clave— FFT, Transformada, Bayes.

***Abstract*— In this article we will see what we mean when we talk about speaker or speaker recognition, how it differs from voice recognition for text. The two types of purposes of speaker recognition, which are identification and verification, the latter being the most similar to voice biometric technology. In addition to these two types of views of a system like this, they are also divided into two depending on whether they are dependent or independent of text, the dependents being the easiest systems to extract parameters when analyzing the speech spectrum. The types of information that can be extracted from a given audio, the implications of this and its difficulty of extraction. The different types of voice analysis, the most prominent those that are based on the spectrum of this, and making use of the Fourier transform (in this case specifically FFT) how we can get parameters and the bases of these methods. Finally, the most appropriate speaker selection mode based on a training bench and an analyzed audio, based on probabilities and Bayes' theorem.**

***Key Word* —FFT, Transformed, Bayes.**

I. INTRODUCCIÓN

La biométrica desde muchos años atrás parecía un cuento de ciencia ficción, algo lejano a nuestro conocimiento. Con el pasar de los años esta tecnología fue viendo la luz poco a poco, todo empezó con el uso de la inteligencia artificial en muchos campos y la evolucion de esta a sistemas mas complejos. Por lo tanto, la biometría ya estaba al alcance. En la actualidad hay diferentes formas de verificar a una persona, todavía existen las herramientas de software que protegen a los dispositivos de intrusos, pero con la introducción de la biométrica estos sistemas se volvieron mas robustos y confiables. La biometría se podría definir como la medición de aspectos biológicos o físicos de los seres, en este caso de los humanos. Ahora existen sistemas de seguridad basados en biometría, como el reconocimiento facial o la huella digital, ya ampliamente usados en smartphones, estos sistemas permiten tener una identidad única a la cual analizar. Pero hay una tercera forma de utilizar esta biométrica en otro aspecto humano, y es la voz. La voz al igual que los otros dos aspectos humanos como la cara y la huella, es única, no hay dos voces iguales, puede que sean muy parecidas pero cada una tiene sus matices o características. Lo que interesa en este articulo es analizar de forma breve como funciona un sistema de reconocimiento de voz, como hace este sistema para saber que persona esta hablando y si es la indicada. Para esto nos vamos a meter en un tema que es muy relacionado con el reconocimiento de voz, pero a la vez difiere mucho de su uso y tecnica (aunque los dos tengas temas centrales como el procesamiento de la voz), este tema es Reconocimiento de Locutor.

II. CONTENIDO

RECONOCIMIENTO DEL LOCUTOR O HABLANTE

El reconocimiento de locutor se puede definir como la extracción de parámetros y características de la voz de alguien

para luego identificarlo. En otras palabras, "*si tenemos un segmento del habla H, y un hipotético locutor L, la tarea de reconocimiento de locutor, se basa en determinar si la locución H fue generada por el locutor L, o dada una cohorte de locutores. Determinar si la locución fue generada por uno de estos o no* "[1]. Para esto se saca un conjunto de patrones del audio correspondiente al locutor y de acuerdo a un conjunto anterior de este guardado se compara, dando una diferencia y tomando una decisión de acuerdo a un umbral, de si pertenece o no al locutor.

El reconocimiento de locutor se caracteriza por dos etapas, el entrenamiento y el reconocimiento. El entrenamiento es el momento en el cual se registran varios locutores por archivos de audio o micrófono, para luego hacer la extracción de caracterizticas y guardarlas en una base de datos, en este punto no se va a centrar mucho en esto, ya que la extracción de las características es de la misma manera tanto para el entrenamiento como para el reconocimiento, es la misma tecnica, lo que se diferencia es las finalidades. Por lo tanto, lo único que podría interesar a grandes rasgos del entrenamiento es como estos datos después de guardados se pueden sacar y comparar, en este caso se podría hacer uso de las redes neuronales para clasificación, pero se tocara el tema un poco mas adelante.

El reconocimiento por otro lado se divide en dos finalidades:

identificación del locutor: En este caso una muestra de voz del locutor es tomada, analizada y comparada con la base de datos antes hecha para sacar un resultado. Para la identificación solo es necesario identificar al locutor, ósea saber que persona esta hablando. Para este contexto de reconocimiento existen dos formas de clasificar estos sistemas, cerrados y abierto. En los sistemas cerrados el locutor existe dentro del conjunto de datos N de entrenamiento, por lo tanto, el numero de posibles decisiones es igual a la dimensión de la población. Por otro lado, en los sistemas abiertos un locutor puede que no exista en el conjunto N de datos, por lo tanto, habría una opción de mas y es de saber si el locutor es uno conocido o desconocido.

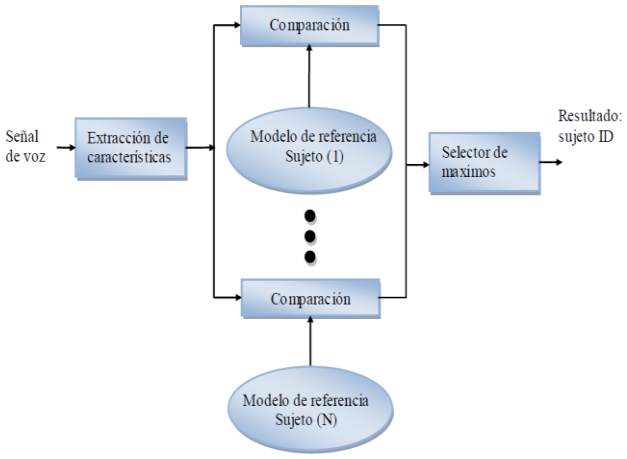

Imagen 1. identificación. Sacada de [1].

Verificación del locutor: Para la verificación, la finalidad es determinar si el locutor es quien dice ser. Para esto el locutor dice la identidad que supuestamente tiene, esta muestra de voz es comparada con el modelo de identidad correspondiente al locutor (donde previamente había una base de datos con datos de la voz de el), para determinar si se acepta o rechaza, el resultado de la comparación se mira de acuerdo con un umbral. Algo para resaltar de este modo, es que, independiente del conjunto de población N, solo hay dos posibles resultados, aceptado o rechazado.

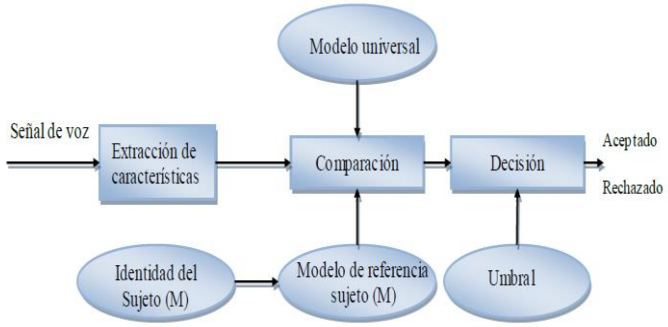

Imagen 2. verificación. sacada de [1].

El sistema de reconocimiento de locutor puede funcionar de dos maneras:

Dependiente de texto: Para este modo es necesario que el locutor provea un audio de su voz diciendo una palabra especifica, que se le podría determinar contraseña. Este es el modo mas simple de reconocimiento de locutor, ya que los patrones se pueden alinear con mas facilidad en la búsqueda con los datos del entrenamiento. El texto puede ser diferente

para cada persona por lo que los datos se pueden diferencia y clasificar mejor.

Independiente de texto: En este caso, el locutor solo cuenta con su voz y un determinado tiempo para ingresar cualquier audio proveniente de su voz, independiente de lo que diga, el sistema debe funcionar. Este es un tipo de sistema mas complejo.

Antes de empezar con el análisis del procesamiento de voz, cabe destacar los niveles de información que se manejan y se tienen en cuenta en este tipo de sistemas.

En la mayoria de conocimiento de estos sistemas, el pilar en el que se basa su funcionamiento es el análisis de señales físicas como son las ondas producidas por la voz, pero hay otros tipos de datos que pueden ayudar en el reconocimiento, estas variables se les llama de alto nivel, ya que son variables que se encuentran en el mas alto nivel del contexto de la voz. Hablamos de aquellas como semántica, dicción, pronunciación, idiosincrasia y mas [1]. El problema de este tipo de variables es que son mas difíciles de sacar, interpretar y colocarlas en un contexto adecuado. Si vemos la imagen 3, podemos apreciar el grado de dificultad de cada tipo de información.

Imagen 3. Peldaños de la información. Sacado de [1]

Espectral: Sale del espectro del sonido de la voz, cada persona tiene su espectro.

Prosódico: Medida de la tensión, acento y entonación. La manera mas fácil de estimarlas es analizando el tono, de la energía y duración. En el lenguaje es cuando hacemos uso de los acentos como silabas tónicas.

Fonético: Reconocer la variabilidad en la pronunciación de un fonema.

Idiolectical: La forma de hablar de acuerdo con el idioma.

Dialógico: Viene de la descripción del dialogo de dos personas, en este caso analizar el dialogo de estas personas y segmentarlo.

ANALISIS DEL PROCESAMIENTO DE VOZ (TECNICAS)

Las técnicas que existen para procesamiento voz son muchas, a lo largo del tiempo se han desarrollado diferentes visiones de como volver la señal de voz en representaciones paramétricas para sacar patrones e identificarlos de forma fácil. Una de las formas mas famosas es por medio del espectro utilizando transformada de Fourier, para este caso es muy apropiado ya que en términos del tiempo la señal no nos da mucha información, diferente es esto en el mundo de la frecuencia. Por esto vamos a ver un método, Análisis por medio de Fourier, del cual se desprenden muchos mas.

Análisis Fourier:

Una de las mas utilizadas soluciones a este problema es por medio del algoritmo rápida transformada de Fourier, aplicando las funciones ventanas para poder analizar tramos de la señal. Para esto se definirá la función ventana como g(t), se definirá un factor de traslación τ. Se toma la señal a analizar, se multiplica con la ventana en la posición adecuada y seguidamente se lleva acabo la transformación frecuencial:

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$$

Ecuación 1. Sacada de [1]

$$S_\tau(\omega) = \hat{f}_g(\tau, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(f(t) g(t-\tau) \right) e^{-j\omega t} dt$$

Ecuación 2. Sacada de [1]

La idea de la función ventana es eliminar al máximo los instantes de tiempo colindantes que se pueden expresar mal en la frecuencia.

definimos $f_g(t,\tau)$ como la función resultante tras el enventanado $f_g(t,\tau) = f(t).g(t-\tau)$

buscaremos que:

$$f_g(t,\tau) \sim \begin{cases} f(t) \text{ pata instantes de tiempo } t \text{ cercanos a } \tau \\ 0 \text{ para instantes de tiempo } t \text{ lejanos a } \tau \end{cases}$$

Imagen 4. Sacada de [1]

Si t=τ es el instante de observación, estaríamos estudiando la distribución de frecuencia alrededor del instante de tiempo de análisis.

Se podría decir que este método es la base de muchos mas como lo es el análisis ceptral, el cual consta del uso de la transformada de Fourier para el espectro de la señal, pero en una escala logarítmica. O el Mel Frequency Ceptral Coefficients, el cual es un complemento para extraer parámetros, donde se vuelve a hacer uso de la función ventana, pero de tipo Hanning o Hamming, y además de la transformada discreta del coseno (DCT).

Otro de los problemas que aparecen al final es como el sistema suelta un resultado de acuerdo a la pieza de audio del locutor, en otras palabras, después de sacar los parameros en el entrenamiento y tener la base de datos, como hace el sistema para decidir la identidad o la verificación. Para esto existen también varios modelos, en este casi mencionaremos uno probabilístico que hace uso del teorema de Bayes. La clasificación de los vectores es basada en verosimilitud, estos vectores, son vectores característicos que salieron del análisis del espectro de voz. Teniendo en cuenta que estos vectores se rigen por una función de distribución de probabilidad. Se tiene:

Asumiendo que los locutores tienen una distribución conocida, con funciones de densidad de probabilidad continuas, entonces la probabilidad de que un vector característico x haya sido generado por el i-ésimo locutor es. Usando la Regla de Bayes, la probabilidad de que, dado un x, el locutor sea el i-ésimo es [2]:

$$p(locutor = i|x) = \frac{p_i(x)p_i}{p(x)} = \frac{p_i(x)}{\sum_{j=1}^{n} p_j(x)}$$

Ecuación 3. Sacada de [2]

P(x) es la probabilidad de que el vector haya sido generado por otro locutor. N es el numero de locutores. $p_i = p_j$.

Acá el problema final es como estimar las densidades de probabilidad $p_i(x)$. Para esto se utiliza la estimación de densidad de probabilidad por medio del método Vecino Mas Cercano.

III. CONCLUSIONES.

El reconocimiento del locutor es otra de las tantas aplicaciones de la inteligencia artificial, y prácticamente es hermana del reconocimiento de voz a texto. Con el tiempo se ha implementado en mas sistemas, como seguridad hablando desde una perspectiva de la biométrica, o simplemente como un sistema de estudio. Es el caso tal de avance de esta tecnología que ahora existen los audios hechos por IA de una persona famosa, por lo que ya es posible falsificar un audio de alguna persona, ya sea de discursos, conversaciones, etc. Este es un paso tanto asombroso como peligroso, ya que se podría llegar a acusar a alguien de algo que nunca dijo.

A pesar de todas estas aplicaciones de esta tecnología, no quita el hecho de que es un tema complejo, hablando técnicamente. Las bases de un sistema de estos se basan en pura matemática, algebra lineal y estadística como probabilidades, además de que toca tener un conocimiento moderado de estos temas para poder entender a fondo como un sistema de estos manipula un audio y saca información por medio del espectro de este. Se podría decir que el reconocimiento del hablante es tan interesante como complejo.

IV. REFERENCIAS.

[1]. Técnicas para reconocimiento automático de locutores-Serie D. Christian, Universidad tecnológica del Bolivar,2007. https://biblioteca.utb.edu.co/notas/tesis/0040351.pdf
[2]. Reconocimiento de Locutor Basado En Procesamiento De Voz-Universidad del Rosario, Argentina. https://www.fceia.unr.edu.ar/prodivoz/speaker_verification.pdf