Reconocimiento de voz en archivos de audio

Speech recognition in audio files

Autor: Steven Medina Gonzalez
IS&C, Universidad Tecnológica de Pereira, Pereira, Colombia
Correo-e: Steven.medina@utp.edu.co

***Resumen*— Este documento presenta un resumen de un problema planteado en clase, acerca de cómo funciona el reconocimiento de voz en archivos de audio. El objetivo del documento es brindar una panorámica general de las técnicas y métodos que están presentas al momento de hacer el debido reconocimiento e identificación de voz de una persona con respecto a un audio, para comprobar que efectivamente es esa persona la autora de dicho audio.**

***Palabras clave*— audio, archivos, voz, técnicas, métodos, reconocimiento, identificación.**

***Abstract*—This document presents a summary of a problem posed in class, about how speech recognition works in audio files. The objective of the document is to provide a general overview of the techniques and methods that are present at the time of making the proper recognition and identification of a person's voice with respect to an audio, to verify that that person is indeed the author of said audio.**

***Key Word*— audio, files, voice, techniques, methods, recognition, identification.**

I. INTRODUCCIÓN

El reconocimiento automático del habla o reconocimiento automático de voz es una disciplina de la inteligencia artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras. El problema que se plantea en un sistema de este tipo es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido, además de esto hay que agregar algo más a este campo y es el reconocimiento de la persona que hizo el audio, reconociéndolo como el autor de dicho archivo.

El reconocimiento de locutores pertenece a la rama de la inteligencia artificial y consiste en la identificación automática de 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 de 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.

II. HISTORIA DEL RECONOCIMIENTO DE VOZ

La historia del reconocimiento de voz empezó en el año de 1870. Alexander Graham Bell quiso desarrollar un dispositivo que capaz de proporcionar la palabra visible para la gente que no escuchara. Bell no tuvo éxito creando este dispositivo, sin embargo, el esfuerzo de esta investigación condujo al desarrollo del teléfono. Más tarde, en los años 30 Tihamer Nemes científico húngaro quiso patentar el desarrollo de una máquina para la transcripción automática de la voz. La petición de Nemes fue negada y a este proyecto lo llamaron poco realista.

Fue hasta 1950, 80 años después del intento de Bell, cuando se hizo el primer esfuerzo para crear la primera máquina de reconocimiento de voz. La investigación fue llevada a los laboratorios de AT&T. El sistema tuvo que ser entrenado para reconocer el discurso de cada locutor individualmente, pero una vez especializada la máquina tenía una exactitud de un 99 por ciento de reconocimiento.

El primer sistema de reconocimiento de voz fue desarrollado en 1952 sobre una computadora analógica que reconocía dígitos del 0 al 9, este sistema era dependiente del locutor. Los experimentos dieron una exactitud de reconocimiento del 98%. Más tarde, en esa misma época, se creó un sistema que reconocía consonantes y vocales.

Durante los 60’s, los investigadores que trabajaban en el área de reconocimiento de voz empezaron a comprender la complejidad del desarrollo de una verdadera aplicación dentro del reconocimiento de voz, y se comenzaron a realizar aplicaciones con vocabularios pequeños, dependientes del locutor y con palabras de flujo discreto. El flujo discreto es la forma como hablan los locutores, es decir, con pequeñas pausas entre palabras y frases. También, durante 1960, la Universidad de Carnegie Mellon e IBM empezaron una investigación en reconocimiento de voz continuo. El impacto de esta investigación se reflejó hasta después de los años 70’s.

Para los 70's, se desarrolló del primer sistema de reconocimiento de voz comercial. Se mejoraron las aplicaciones de los sistemas dependiente del locutor que requerían una entrada discreta y tenía un vocabulario pequeño.

Por otra parte, la Advanced Research Projects Agency (ARPA) de la Sección americana de Defensa se mostró interesada en la investigación de reconocimiento de voz. ARPA comenzó investigaciones enfocándose al habla continua y usando vocabularios más extensos. También se mejoró la tecnología de reconocimiento para palabras aisladas y continuas. En esta misma época se desarrollaron técnicas para el reconocimiento de voz como tiem warping, modelado probabilístico y el algoritmo de retro propagación.

Durante los 80's el reconocimiento de voz se favoreció por tres factores: el crecimiento de computadoras personales, el apoyo de ARPA y los costos reducidos de aplicaciones comerciales. El mayor interés durante este periodo de tiempo era el desarrollo de vocabularios grandes. En 1985 un vocabulario de 100 palabras era considerado grande. Sin embargo, en 1986 hubo uno de 20,000 palabras. También durante esta época hubo grandes avances tecnológicos, ya que se cambió del enfoque basado en reconocimiento de patrones a métodos de modelado probabilísticos, como los Modelos Ocultos de Markov (HMM).
Para los 90's los costos de las aplicaciones de reconocimiento de voz continuaron decreciendo y los vocabularios extensos comenzaron a ser normales. También las aplicaciones independientes del locutor y de flujo continuo (lo contrario al flujo discreto, es decir, en el habla no hay pausas significantes) comenzaron a ser más comunes.
En esta subsección se ha proporcionado una breve introducción de los sistemas de reconocimiento de voz y su historia. A continuación, se describen las principales características acústicas del habla.

III. VERIFICACIÓN VS IDENTIFICACIÓN

Los dos campos de aplicación más importantes del reconocimiento de locutores son la verificación y la identificación de hablantes. Si el locutor afirma tener una determinada identidad y el sistema debe corroborarla, el sistema está realizando verificación de locutores. Si en cambio el sistema sólo recibe características de una voz y debe determinar su identidad, por ej. dentro de un conjunto de posibles identidades, estamos en ese caso ante un sistema de identificación.

En la verificación de locutores el sistema de reconocimiento verifica si las características extraídas de la voz de un locutor se corresponden con la identidad que afirma tener el mismo. 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. La verificación de locutores se utiliza típicamente en seguridad (por ej. para dar acceso a una puerta).

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.

En aplicaciones forenses (por ej. en investigaciones policiales o evaluación de evidencias en la justicia), 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.

IV. ADQUISICIÓN DE DATOS

La adquisición de datos es esencial tanto para la parte de entrenamiento como para la de test. Para poder introducir locutores al sistema es necesario un transductor acústico-eléctrico, ya que la voz se propaga en forma de ondas y para poder extraer características es necesario transformar la presión sonora en una señal eléctrica y así poder proceder a su digitalización.

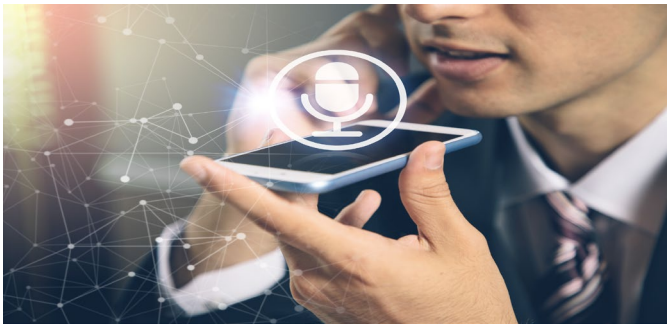

Fig 1. Referencia de toma de datos

El tipo de micrófono, la frecuencia de muestreo y la cuantización realizada en la captación del audio deberá adecuarse al ancho de banda de la voz y sus características. Hay factores externos al locutor como la elección de los parámetros anteriores, la relación señal ruido (SNR) de las muestras grabadas o la utilización de micrófonos con diferentes curvas de respuesta frecuencial que pueden influir negativamente en el resultado.

V. EXTRACCIÓN DE CARACTERÍSTICAS

Una vez digitalizado, el audio se procesa para extraer el listado de características elegidas, las cuales se llaman descriptores de audio. Estos descriptores contienen las características acústicas de la señal que utilizará el clasificador para compararlos con el listado almacenado en la base de datos. Las características a analizar pueden ser diversas, pero se suelen utilizar los descriptores de audio de bajo nivel

debido a la naturaleza de la fuente. Estos descriptores presentan un bajo nivel de abstracción y se limitan a describir características espectrales, paramétricas y temporales de la señal de audio.

Para poder asociar las características de los descriptores a los archivos de audio correspondientes se utilizan los metadatos, datos sobre datos. Uno de los estándar utilizados para esta tarea es el estándar MPEG-7, el cual permite la gestión de estos metadatos, facilitando así el acceso a esta información en el momento de la búsqueda.

Parámetros	Rango
Modo de hablar	Palabras aisladas o habla continua
Estilo del habla	Voz de lectura o voz espontánea
Aislamiento	Dependiente del locutor o Independiente del locutor
Vocabulario	Pequeño (<20 palabras) o grande (>20,000 palabras)
Modelo del lenguaje	Estados finitos o dependiente del contexto
Perplejidad	Pequeña (<10) o larga (>100)
Reducción del ruido en el habla	Alta (>30 dB) o baja (<10 dB)

Fig 2. Parámetros que caracterizan a un Sistema de reconocimiento de voz

Existen diferentes niveles en los que la identidad del hablante se encuentra en la señal de voz hablante se encuentra en la señal de voz.
Cuando reconocemos a alguien por la voz tenemos en cuenta:

- Su timbre
- Su uso de los sonidos Su uso de los sonidos
- Su forma de entonar

Esa combinación es dependiente del locutor a reconocer

Las diferentes características de la voz se agrupan en niveles (lingüística):

- Fonético: utilización de Fonético: utilización de diferentes sonidos, pronunciación, etc.
- Prosódico: entonación particular, variación de energía, pausas entre frases o palabras, etc.
- Espectral: configuración (resonancia) tracto vocal, coarticulación, nasalidad, etc.

Fig 3. Referencias características de voz distintas.

VI. CLASIFICACIÓN

El módulo clasificador tiene acceso tanto a la parte de entrenamiento como a la de test. Este módulo hace de puente entre ambas partes encargándose de comparar los vectores de características a buscar con los vectores de los modelos de locutor que contiene la base de datos. Su tarea computacional consiste en encontrar coincidencias y como resultado extrae una serie de probabilidades de los locutores en la base de datos susceptibles de ser el buscado. La decisión puede ser diferente dependiendo de la configuración del sistema.

- Sistema cerrado: Un sistema cerrado da por supuesto que el locutor que se quiere identificar se encuentra ya almacenado en la base de datos. El locutor con más probabilidades a la salida del clasificador, que comparte más características con el locutor a buscar, será la salida resultante del sistema.

- Sistema abierto: Un sistema abierto es más complejo, ya que el locutor que se quiere identificar no está necesariamente en la base de datos. El clasificador debe tener en cuenta no sólo la más alta probabilidad, sino que también debe establecer si la semejanza es suficiente para dar un positivo. Si las probabilidades de un modelo de locutor se consideran suficientes como para suponer una coincidencia se presenta al candidato como resultado de la búsqueda, en caso contrario la salida es "locutor desconocido".

VII. 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.

Campo	Ejemplos
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
Refuerzo de la ley	Comprobación de que se cumple la libertad condicional
Forense	Identificación de personas a través de grabaciones para validar pruebas

Fig 4. Posibles aplicaciones de esta rama de la inteligencia artificial.

VIII. CONCLUSIÓN

Apartide toda la información y fuentes referenciadas en este documento, podemos ver que este campo ha sido muy importante en bastantes temas de la actualidad, teniendo paso desde la investigación forense para identificar posibles sospechosos hasta la seguridad en equipos propios o de empresas, en todo caso hay muchas posibilidades donde hay oportunidad para implementar esta tecnología y mejorar los sistemas.

REFERENCIAS

Referencias en la Web:

[1] https://es.wikipedia.org/wiki/Reconocimiento_de_locutores

[2] http://garciaargos.com/descargas/apuntes/posgrado/Primer-Semestre/Reconocimiento-Biometrico/2008_Master_UAM_Locutor_v4.pdf

[3] http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ahuactzin_l_a/capitulo1.pdf

[4] https://www.researchgate.net/publication/39697457_Redes_neuronales_en_reconocimiento_de_locutor

[5] https://www.fceia.unr.edu.ar/prodivoz/speaker_verification.pdf