Computación Blanda

Soft Computing

Autor: Juan Camilo Varón
IS&C, Universidad Tecnológica de Pereira, Pereira, Colombia
Correo-e: juan.varon@utp.edu.co

Un sistema de reconocimiento de hablantes se resume como aquel capaz de reconocer automáticamente qué persona está hablando de entre las personas pertenecientes a una base de datos de señales de audio previamente clasificadas. Para construir estos sistemas se utiliza la información conocida sobre el tracto vocal de un individuo y las propiedades físicas del sonido.

Estos sistemas se pueden utilizar para innumerables aplicaciones y en numerosos ámbitos. A continuación algunos ejemplos:

- ***Seguridad: Autenticación o prevención de fraude.***
- ***Comercio: Confirmación de compras a distancia.***
- ***Educación: Nuevas enseñanzas o métodos de aprendizaje para personas discapacitadas.***

Palabras clave: Reconocimiento, hablante, locutor, sistema, Cepstrum.

***Abstract*— A speaker recognition system is a process that recognizes the person who is talking among those stored in a database of audio signals previously classified. The system uses information about the vocal tract and the physical properties of sound.**

These systems can be used for a lot of applications and a lot of fields. Some examples:

- **Security: Authentication or fraud prevention.**
- **Trade: Confirmation about distance shopping.**
- **Education: New methods of teaching or disabled people.**

***Key Word*— Recognition, speaker, system, Cepstrum.**

I. INTRODUCCIÓN

La inteligencia artificial fue introducida en el estudio científico en el año 1950 por el inglés Alan Turing, provocando un interés mundial con su pregunta "Can machines think?". Numerosos científicos comenzaron sus investigaciones orientadas a crear modelos y teorías que dieran explicación al funcionamiento de la inteligencia. En la actualidad los estudios de inteligencia artificial se basan en el desarrollo de sistemas de procesado de datos que imitan el comportamiento de la mente humana, con sistemas de decisión tras un aprendizaje previo [INT].

Si bien un gran número de investigadores buscan respuestas en el estudio psicológico, la gran mayoría se apoya en la física para conseguir teorías aproximadas y asimilar sus sistemas al comportamiento humano (el ojo, el oído, ondas, frecuencias, etc.).

La voz humana es definida por la Real Academia de la Lengua como "Sonido que el aire expelido de los pulmones produce al salir de la laringe, haciendo que vibren las cuerdas vocales" o "Calidad, timbre o intensidad de este sonido" [RAE]. El habla consiste en combinar las unidades fónicas (fonemas) que la voz humana es capaz de generar para formar una lengua. Estos símbolos no son idénticos de un individuo a otro, pero poseen características comunes que los hacen descifrables dentro de una lengua o dialecto.

DISEÑO TÉCNICO

Hay numerosos algoritmos con los que se podría construir un sistema de reconocimiento de hablante. Las variables aparecen desde la elección del lenguaje de programación hasta el tipo de audio, el procesado del mismo, las características estudiadas, el algoritmo de comparación o la distancia utilizada.

Se detallará a continuación el diseño de la solución así como la elección de todos los parámetros citados anteriormente.

TEORÍA FUNDAMENTADA EN LA VOZ

El ser humano es capaz de generar voluntariamente ondas de presión acústica a partir de movimientos de la estructura anatómica del sistema fonador humano. La generación de voz comienza en el cerebro con la conceptualización de la idea que se quiere transmitir, la cual se asocia a una estructura lingüística seleccionando las palabras adecuadas y ordenándolas de acuerdo con unas reglas gramaticales. A continuación el cerebro produce impulsos nerviosos que mueven los órganos vocales para producir los sonidos. Los órganos involucrados son las cuerdas vocales, el paladar, la lengua, los dientes, los labios y la mandíbula.

La frecuencia de cada sonido depende de varios factores, pero principalmente del tamaño y la masa de las cuerdas vocales y de la tensión que se les aplique, lo que hará que el aire salga de los pulmones a una velocidad u otra:

- Mayor tamaño: Menor frecuencia (graves)
- Mayor tensión: Mayor frecuencia (agudos)

La zona que incluye la cavidad faríngea, oral y nasal junto a los elementos articulatorios se denomina cavidad supraglótica mientras que los espacios por debajo de la laringe como la tráquea, los bronquios y los pulmones conforman las cavidades infraglóticas.

El amplio margen de sonidos es posible gracias a que algunos de los elementos de la cavidad supraglótica se controlan a voluntad. La faringe y las cavidades nasal, oral y labial realizan un filtrado modificando el espectro, actuando como resonadores acústicos que enfatizan determinadas bandas de frecuencia reforzando la amplitud de grupos de armónicos situados alrededor de una determinada frecuencia.

Todos los sonidos tienen por tanto variables dependientes del hablante, como el tamaño de las cuerdas vocales, pero también variables en común con otros hablantes, como el rango de frecuencia de un determinado fonema. El conjunto de estas variables hará posible cualquier estudio sobre voz humana.

IMPLEMENTACION

Para implementar el sistema de reconocimiento de hablantes se podría haber elegido entre una amplia variedad de algoritmos. En este caso se ha elegido una combinación que no tiene por qué ser óptima, ya que para hacer esta afirmación se necesita un estudio riguroso de todas las demás, pero que tiene cierto sentido lógico a priori y da unos resultados acordes al objetivo buscado. Se ha seguido la estructura típica de los trabajos de sistemas clasificatorios con reconocimiento de patrones, apareciendo los cuatro bloques principales: procesado, extracción de características, algoritmo de clasificación y toma de decisión.

Se diferencian dos etapas bien definidas en el trabajo de investigación: la etapa de training o entrenamiento y la etapa de testing o clasificación.

La etapa de training consiste en introducir en el programa una serie de audios de los cuales se conoce su locutor original. Es necesario aplicar a estos audios el mismo procesado que se aplicará posteriormente a los audios en la etapa de testing para adecuarlos a las condiciones idóneas para trabajar (misma frecuencia de muestreo, reducción de ruido y eliminación de silencios). Una vez hecho esto, se procede a extraer las características. Trabajaremos con coeficientes Cepstrum, una serie de coeficientes numéricos que aportan información sobre la señal basados en la percepción auditiva humana, derivados de la transformada de Fourier por tramas (o ventanas) de sonido. Cada audio de entrenamiento proporciona una matriz de coeficientes que se irán concatenando verticalmente hasta formar la matriz de entrenamiento, que será la que contenga los coeficientes cepstrales de todos los audios de train.

La etapa de clasificación es la que se encarga de analizar el audio desconocido para asociarlo finalmente a uno de los locutores con los que se ha entrenado el programa. El procesado es el mismo que en la anterior etapa, exceptuando que la matriz de clasificación sólo contendrá los coeficientes Cepstrum del audio a estudiar.

Una vez se tienen las dos matrices, es importante normalizarlas para evitar futuros problemas por los distintos rangos de las variables. Si los rangos y varianzas son semejantes, el efecto será muy reducido pero si no es así, puede haber grandes desbalances.

El siguiente paso es utilizar un algoritmo que determine a qué locutor de entrenamiento se parece más cada trama de coeficientes del locutor a clasificar. De nuevo hay numerosos algoritmos para esta función; el que va a utilizarse en este caso es el algoritmo K-NN o de N vecinos cercanos. Este algoritmo asocia 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.

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.

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.

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.

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.

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.

REFERENCIAS

Referencias en la Web:

[1]
https://www.eldiario.es/tecnologia/diario-turing/reconocimiento-voz-biometria_1_5143374.html

[2]
http://physionet.cps.unizar.es/~eduardo/investigacion/voz/rahframe.html

[3]
http://www.scielo.org.mx/pdf/cys/v9n3/v9n3a7.pdf