Reconocimiento Vocal Por Sistemas Computacionales

Voice Recognition by Computational Systems

Autor: Santiago Torres Vasquez

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

Correo-e: santiago.torres@utp.edu.co

*Resumen*— El documento presente consta de un resumen de las características del reconocimiento vocal, así como distintos métodos implementados para su cumplimiento. El objetivo del documento es brindar una perspectiva general de los detalles y dificultades, así como soluciones, que nacen con el desarrollo de este paradigma. El reconocimiento vocal consta de identificar las distintas características presentes en las ondas sonoras generadas por un individuo, tomando en cuenta los factores que causan una identidad fonética, y su uso para comparar dichas identidades para reconocer y/o verificar la identidad de un individuo.

*Palabras clave*— voz, identidad, ondas sonoras, características, reconocimiento.

*Abstract*— This document consists of a summary of the characteristics of voice recognition, as well as different methods implemented for its fulfillment. The objective of the document is to provide an overview of the details and difficulties, as well as solutions, that arise with the development of this paradigm. Voice recognition consists of identifying the different characteristics present in the sound waves generated by an individual, taking into account the factors that cause a phonetic identity, and its use to compare said identities to recognize and / or verify the identity of an individual.

*Key Words*— voice, identity, sound waves, characteristics, recognition.

1. INTRODUCCIÓN

El reconocimiento de voz es el proceso de reconocer automáticamente quién está hablando mediante el uso de información específica del locutor incluida en las ondas del habla para verificar la identidad de este. Los servicios aplicables incluyen marcación por voz, distintas transacciones por teléfono como compras y acciones bancarias, control y acceso a sistemas de seguridad, entre otros. Otra aplicación importante de la tecnología de reconocimiento de hablantes es como herramienta forense.

Las características fisiológicas del locutor están directamente relacionadas con su identidad fonética. Las características del tracto vocal y las características supra segmentarias del habla crean en conjunto una identidad vocal bastante única. Las mediciones espectrales a corto plazo más utilizadas son los coeficientes cepstrales y sus coeficientes de regresión. En cuanto a los coeficientes de regresión, típicamente, los coeficientes de primer y segundo orden, es decir, las derivadas de las funciones de tiempo de los coeficientes cepstrales, se extraen en cada período de sección para representar la dinámica espectral. Estos coeficientes de regresión se denominan respectivamente coeficientes delta-cepstral y delta-delta-cepstral.

1. PRINCIPIOS

El reconocimiento de locutor se puede clasificar en identificación de locutor y verificación de locutor, de forma similar al proceso de identificación y verificación en un sistema de seguridad.

La identificación del locutor es el proceso de determinar de cuál de las personas registradas en la base de datos proviene una determinada expresión fonética. La verificación del locutor es el proceso de aceptar o rechazar la identidad que dicho locutor reclama como suya. La mayoría de las aplicaciones en las que se utiliza la voz para confirmar la identidad de un locutor se clasifican como verificación del locutor.

Aunque las tareas luzcan similares, existen diferencias fundamentales entre ellas: la identificación tiene como parámetro de salida una identidad, siendo aquella que mejor coincida con el locutor tomando una población muestral (La base de datos), a diferencia de la verificación, en donde se devuelve un parámetro booleano; un sí o un no, dependiendo de si la muestra del locutor coincide con la muestra de la persona que reclama ser, siendo esta una población muestral de uno.

En la tarea de identificación del locutor, se analiza una muestra fonética de un locutor desconocido y se compara con muestras y modelos fonéticos de personas registradas. El locutor desconocido se identifica como la persona cuyo modelo se adapta mejor a la muestra de entrada. En la verificación del locutor, un locutor desconocido reclama una identidad, y la muestra de este locutor desconocido se compara con un modelo de la persona cuya identidad se reclama. Si la coincidencia es lo suficientemente buena, es decir, por encima de un umbral, se acepta la declaración de identidad. Un umbral alto dificulta que los impostores sean aceptados por el sistema, pero con el riesgo de rechazar falsamente a los usuarios válidos. Por el contrario, un umbral bajo permite que los usuarios válidos sean aceptados de forma coherente, pero con el riesgo de aceptar impostores. Para establecer el umbral en el nivel deseado de rechazo del cliente (falso rechazo) y aceptación del impostor (falsa aceptación), se necesitan datos que muestren las distribuciones de las puntuaciones del cliente y del impostor.

Para evaluar la eficacia de los sistemas de verificación del locutor se puede utilizar la curva ROC: la curva de características operativas del receptor, adoptada desde el campo de la psicofísica, la cual se obtiene asignando dos probabilidades, la probabilidad de aceptación correcta (Tasa de falso rechazo) y la probabilidad de aceptación incorrecta (tasa de falsa aceptación), a los ejes vertical y horizontal respectivamente, y variando el umbral de decisión. Por otro lado, la tasa de error equivalente es una medida general comúnmente aceptada del rendimiento del sistema. Corresponde al umbral en el que la tasa de aceptación falsa es igual a la tasa de rechazo falso.

Los métodos de reconocimiento de locutor también se pueden dividir entre métodos dependientes del texto e independientes del texto. Los primeros requieren que el locutor proporcione enunciados de palabras u oraciones clave, las cuales también fueron usadas al momento de registrar nuevos individuos y entrenar el sistema, mientras que los segundos no requieren que se pronuncie un texto específico.

1. MÉTODOS DEPENDIENTES DE TEXTO

Los métodos dependientes de texto generalmente se basan en técnicas de coincidencia de secuencia plantilla-modelo en las que los ejes de tiempo de una muestra de voz de entrada y las plantillas de referencia o modelos de referencia de los hablantes registrados se alinean, y las similitudes entre ellos se acumulan desde el principio hasta el final del enunciado. Dado que este método puede explotar directamente la individualidad de la voz asociada con cada fonema o sílaba, generalmente logra un mayor rendimiento de reconocimiento que el método independiente del texto.

Los métodos de reconocimiento de locutor dependientes de texto se pueden clasificar en métodos basados ​​en distorsión de tiempo dinámica (DTW) o modelo oculto de Markov (HMM).

En la primera categoría, cada enunciado está representado por una secuencia de vectores de características, generalmente vectores de características espectrales a corto plazo, y la variación de tiempo de ensayo a ensayo de los enunciados del mismo texto se normaliza alineando la secuencia de vectores de características analizada de una prueba con la secuencia de vectores de características de una plantilla utilizando un algoritmo DTW. La distancia total entre el enunciado de prueba y la plantilla se utiliza para tomar la decisión de reconocimiento. En el caso de utilizar distintas plantillas para representar la variación espectral, las distancias entre la expresión de prueba y las plantillas se promedian y luego se utilizan para tomar la decisión.

Un HMM puede modelar de manera eficiente la variación estadística en características espectrales. Por lo tanto, los métodos basados ​​en HMM han logrado precisiones de reconocimiento significativamente mejores que los métodos basados ​​en DTW.

1. MÉTODOS INDEPENDIENTES DE TEXTO

En el enfoque del reconocimiento del locutor independiente de texto, no se puede evaluar la concordancia de una muestra en base a la comparación de enunciados. Dado que es imposible modelar o hacer coincidir los eventos del habla a nivel de palabra u oración, se han investigado los siguientes cuatro tipos de métodos: métodos basados en estadísticas a largo plazo, métodos de cuantificación de vectores (VQ), métodos de HMM ergódicos, y métodos basados en reconocimiento del habla.

Los métodos basados en estadísticas a largo plazo se basan en el uso de estadísticas de muestra a largo plazo de varias características espectrales, como la media y la varianza de las características espectrales en una serie de expresiones. Los promedios espectrales a largo plazo son condensaciones extremas de las características espectrales de las expresiones de un hablante y, como tales, carecen del poder de discriminación de las secuencias de características espectrales a corto plazo utilizadas como modelos en métodos dependientes del texto.

Por otro lado, los métodos VQ consisten en utilizar libros de código de cuantificación de vectores, que constan de un pequeño número de vectores de características representativos, se utilizan como un medio eficaz para caracterizar características específicas del hablante. En la etapa de reconocimiento, un enunciado de entrada se cuantifica por vector utilizando el libro de códigos de cada locutor de referencia; la distorsión acumulada en todo el enunciado de entrada se utiliza para realizar la determinación de reconocimiento.

En los métodos de HMM ergódicos, la estructura básica es la misma que la del método basado en VQ, reemplazando los libros por un HMM diseñado de forma específica. En una larga escala de tiempo, la variación temporal en los parámetros de la señal de voz está representada por transiciones estocásticas de Markov entre estados. Este método utiliza un HMM ergódico de múltiples estados, permitiendo todas las posibles transiciones entre estados, para clasificar los segmentos de voz en una de las categorías fonéticas generales correspondientes a los estados del HMM. Las categorías obtenidas automáticamente se caracterizan a menudo por la fuerza de la voz, sus silencios, la cantidad de fricación, etc.

Los métodos basados ​​en VQ y HMM pueden considerarse métodos que utilizan detalles del locutor dependientes de la clase de fonema, contenidas en características espectrales a corto plazo a través del reconocimiento implícito de la clase de fonema. En otras palabras, las clases de fonemas y los locutores se reconocen simultáneamente en estos métodos. Por otro lado, en los métodos basados ​​en el reconocimiento del habla, los fonemas y las clases de fonemas se reconocen explícitamente, para luego ser comparadas con modelos del locutor o plantillas correspondientes a ese fonema o clase.

Se ha investigado el uso de HMM predictivos ergódicos de cinco estados para una categorización fonética amplia. En el método de reconocimiento del habla, una vez que se han identificado las secciones que pertenecen a categorías fonéticas, las cuales tienen distintos pesos, y se realiza la selección de características. Cada umbral de verificación de características fue definido durante la fase de entrenamiento, en la cual también se generaron plantillas de referencia, las cuales son usadas durante la fase de verificación para comparar cada categoría y generar una puntuación. La puntuación de verificación final es una combinación lineal ponderada de las puntuaciones de cada categoría. Los pesos se eligen para reflejar la eficacia de categorías particulares de fonemas para discriminar entre hablantes y se ajustan para maximizar el rendimiento de la verificación. Los resultados experimentales mostraron que la precisión de la verificación se puede mejorar considerablemente mediante este método de combinación lineal ponderada dependiente de la categoría.

REFERENCIAS

Referencias en la web

[1]

<http://www.scholarpedia.org/article/Speaker_recognition>