Reconocimiento automático del habla

Helenca Duxans Barrobés
Marta Ruiz Costa-jussà

PID_00188070

UOC
Universitat Oberta de Catalunya

Índice

Introducción

En este módulo introducimos el concepto de reconocimiento automático del habla y las aplicaciones que tiene. A continuación, veremos la estructura genérica más utilizada por los sistemas de reconocimiento automático del habla y detallaremos cada uno de los módulos que lo forman. Finalmente, presentaremos cómo se evalúa el funcionamiento de los reconocedores, es decir, cómo se mide el nivel de calidad de las transcripciones que aquellos nos proporcionan y cómo se pueden mejorar por medio de técnicas de adaptación.

¿Por qué se debe estudiar el reconocimiento del habla?

Hablar es el medio más natural para las personas a la hora de comunicarnos. Para interactuar del mismo modo con el mundo digital que nos rodea, lo primero que necesitamos es que las máquinas (el ordenador, el teléfono, el coche, etc.) nos entiendan. El reconocimiento del habla proporciona las herramientas necesarias para transformar la voz en conceptos que después puedan utilizar las máquinas para emprender acciones.

Objetivos

Al acabar de trabajar este módulo, deberéis ser capaces de lo siguiente:

1. Explicar el funcionamiento de un reconocedor a alto nivel.

2. Identificar los módulos principales de los reconocedores automáticos del habla, junto a las funciones que estos tienen.

3. Saber para qué se utilizan los modelos ocultos de Markov en el reconocimiento del habla.

4. Distinguir los conceptos de diccionario y el modelo de lenguaje.

5. Evaluar una transcripción automática proporcionada por un reconocedor.

1. Introducción al reconocimiento automático del habla

El reconocimiento automático del habla, denominado también *conversión de voz a texto*, conocido por las siglas ASR [*automatic speech recognition* (reconocedor automático del habla)], consiste en transcribir el contenido de una señal de voz sin intervención humana.

Los reconocedores del habla se clasifican según diferentes criterios:

- **El estilo del habla** o el tipo de habla que pueden reconocer; por ejemplo, habla aislada (palabras separadas por pausas), habla continua o natural (frases) y habla espontánea (discurso, incluyendo repeticiones, interjecciones, tonos, etc.).

- **La dependencia del locutor**. Los reconocedores dependientes de los locutores aprenden las características de la voz de un usuario o un conjunto de usuarios concretos para mejorar la transcripción, y por lo tanto funcionan de manera óptima para estos usuarios. En cambio, los reconocedores independientes del locutor están diseñados para tener unas prestaciones similares sea quien sea el usuario, es decir, no están limitados a un conjunto de usuarios conocidos.

- **El tipo de canal**. Los reconocedores pueden ser construidos para trabajar con audio proveniente de micrófono, telefonía fija o telefonía móvil.

- **Otros**. Otros criterios de clasificación de un reconocedor serían si este es multiidioma o la dimensión del vocabulario que puede reconocer.

Por ejemplo, los sistemas de atención telefónica que piden que respondamos a preguntas con un *sí* o un *no*, diciendo el nombre de un departamento, etc., antes de dirigir la llamada adecuadamente a una persona real se clasifican como reconocedores de habla aislada, independientes del locutor y diseñados para telefonía tanto fija como móvil.

Reflexión

¿Creéis que los reconocedores independientes del locutor son igual de fiables para todos los usuarios? La respuesta es no. Cada persona tiene una manera de hablar única.

2. Aplicaciones de los reconocedores automáticos del habla

Desde la aparición del primer prototipo de ASR a mediados del siglo XX, la tecnología del reconocimiento automático del habla ha avanzado bastante, ya que hoy en día existe en el mercado un abanico amplio de aplicaciones comerciales.

Uno de los primeros ámbitos en los que se desplegó esta tecnología fue en la **atención telefónica**, como los servicios de información, de notificación de averías o de concertar cita. En este tipo de servicios, el sistema hace que los usuarios sigan un diálogo dirigido[1], con el objetivo de obtener la información necesaria para tomar una acción (por ejemplo, proporcionar la información preguntada, dirigir la llamada hacia el operador adecuado, etc.).

(1)Del tipo: "Diga el nombre de la ciudad de la que quiere conocer la predicción meteorológica".

Otro ámbito en el que el uso de los ASR está muy extendido es el de las aplicaciones de **mando** y **control**. Los primeros marcadores vocales, que tienen su origen en los teléfonos móviles, han evolucionado hasta el punto de poder controlar dispositivos en situaciones en las que el uso de otro medio no sea seguro (por ejemplo, por el control de la radio o el GPS dentro del coche) o cómodo (por ejemplo, en aplicaciones de domótica).

Para vocabularios muy concretos, como es el del campo médico, la **transcripción de notas** o informes se utiliza como herramienta de trabajo, gracias a la reducción de tiempo que proporciona al equipo médico. Los sistemas de **dictado**, adaptados a un solo usuario pero sin restricción de vocabulario, tienen una penetración discreta en el mercado doméstico, dado que para que funcionen correctamente requieren un entrenamiento en forma de retroacción o *feedback* del usuario, sobre todo en las primeras etapas de utilización.

El cambio de uso de los teléfonos móviles debido a las conexiones de datos asequibles económicamente también ha abierto nuevas aplicaciones paralelas ASR, como las **búsquedas** automáticas por medio de la voz (por ejemplo, acceso a Google). Aun así, la **indexación de audio**, que permite encontrar información sobre una canción a partir de su reproducción o sobre el tema de un discurso, es una aplicación emergente de los ASR, a pesar de que el despliegue que tiene en el mundo comercial está limitado por el estado de la tecnología actual.

Existen otras aplicaciones, como el **subtitulado automático** o la **transcripción de reuniones**, que –a pesar de ser aplicaciones muy interesantes de los ASR– todavía no se pueden llevar a cabo por las altas tasas de error que tiene la tecnología actual para dominios abiertos.

Sin embargo, los ASR son una ayuda muy valiosa para realizar **transcripciones semiautomáticas**, es decir, transcripciones automáticas revisadas por una persona. Un ejemplo de este tipo de aplicaciones es la transcripción de mensajes telefónicos en SMS.

3. Funcionamiento básico de los reconocedores

Los reconocedores automáticos del habla tienen como entrada un audio y como salida el texto (transcripción) correspondiente. La mayoría de los ASR comparten el diagrama funcional que se muestra en la figura 1:

Diagrama funcional de los reconocedores automáticos del habla

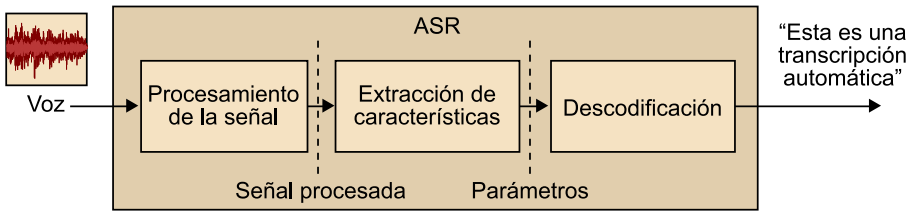

Figura 1. Diagrama funcional de los reconocedores automáticos del habla (ASR)

En el módulo de **procesamiento de la señal**, el audio de entrada se transforma para facilitar el reconocimiento. Entre las operaciones habituales que se efectúan en este módulo están las técnicas de reducción de ruido (ecualización de canal, cancelación de eco, etc.) y la detección de voz y ruido. La detección de voz y ruido consiste en segmentar el audio de entrada y clasificar cada segmento en dos categorías, según si se considera que es voz o ruido. De este modo, se envían únicamente los segmentos de audio correspondientes a voz al módulo de extracción de características y al módulo descodificador.

En el módulo de **extracción de características**, el audio correspondiente a voz llega en segmentos de poca longitud (aproximadamente, milisegundos), y para cada segmento se obtiene un conjunto de parámetros espectrales, denominado *vector de características*. La salida de este módulo es una secuencia temporal (u ordenada) de vectores de características. Por lo tanto, en la salida hay tantos vectores de características como segmentos de voz en la señal de audio original. En el apartado "El módulo de extracción de características" encontraréis una descripción más detallada de esto.

El objetivo del módulo de descodificación es obtener, a partir de la secuencia de vector de características, la transcripción (texto) que con más probabilidad se ha pronunciado. En este proceso intervienen diferentes fuentes de información: información acústica (por medio de los modelos acústicos) e información lingüística (por medio de los diccionarios y modelos de lenguaje o gramáticas). En el apartado "El módulo de descodificación" explicaremos cada uno de los componentes de la descodificación.

Los ASR transforman la voz en texto.

Los ASR están formados por tres módulos: el módulo de procesamiento de la señal, el extractor de características y el módulo de descodificación.

El módulo de procesamiento de la señal limpia el audio de entrada, lo segmenta e identifica los segmentos en los que hay voz.

El extractor de características transforma los segmentos de audio identificados como voz por el módulo de procesamiento de la señal en un conjunto de vectores de características que contienen la información necesaria para reconocerlos.

El módulo de descodificación transforma los vectores de características en texto.

4. El módulo de extracción de características

La voz es una señal continua y variante en el tiempo y, por lo tanto, es difícil de analizar en el dominio temporal. Para convertir la voz en texto, la señal de entrada se transforma en una secuencia discreta de parámetros, concretamente en una secuencia de vectores de características. El módulo de extracción de características es el encargado de extraer estos vectores.

Módulo de extracción de características

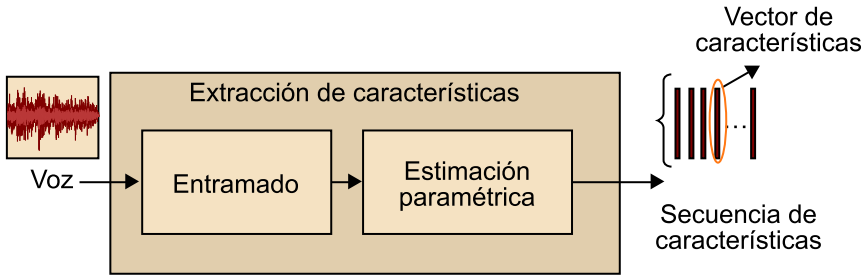

Figura 2. Diagrama de bloques del módulo de extracción de características

Para usar los vectores de características para el reconocimiento, estos vectores han de tener las propiedades siguientes:

- Cada vector de características debe representar un segmento temporal de la señal de voz. Por lo tanto, la señal de voz se ha de dividir en segmentos que se puedan considerar estacionarios (con características estables). El proceso de obtención de segmentos considerados estacionarios se denomina **entramado**. Este proceso se puede considerar como el último del módulo de procesamiento de la señal o como el primero del módulo de extracción de características.

- El vector de características debe contener toda la información necesaria para el reconocimiento del habla de la manera más compacta posible. Idealmente, el vector de características solo ha de contener la información relevante para el reconocimiento y eliminar el resto de la información complementaria que hay de manera intrínseca en el audio (particularidades de pronunciación de cada locutor, ruido de ambiente, etc.). El proceso de obtención de características de cada segmento se denomina **estimación paramétrica**.

4.1. Entramado

Para hacer el **entramado**, en cada instante de análisis, la señal de entrada se multiplica por otra señal, denominada *ventana*, que es diferente de 0 solo en un intervalo temporal. El resultado de cada una de estas multiplicaciones es una señal de duración limitada, que se denomina *tramo* o *segmento*. La figura 3 ilustra este proceso y se muestran tres tramos obtenidos por los instantes de análisis $\{t_i, t_{i+1}, t_{i+2}\}$.

Frame

Habitualmente se utiliza el término inglés *frame* para referirse a un tramo o segmento de la extracción de características.

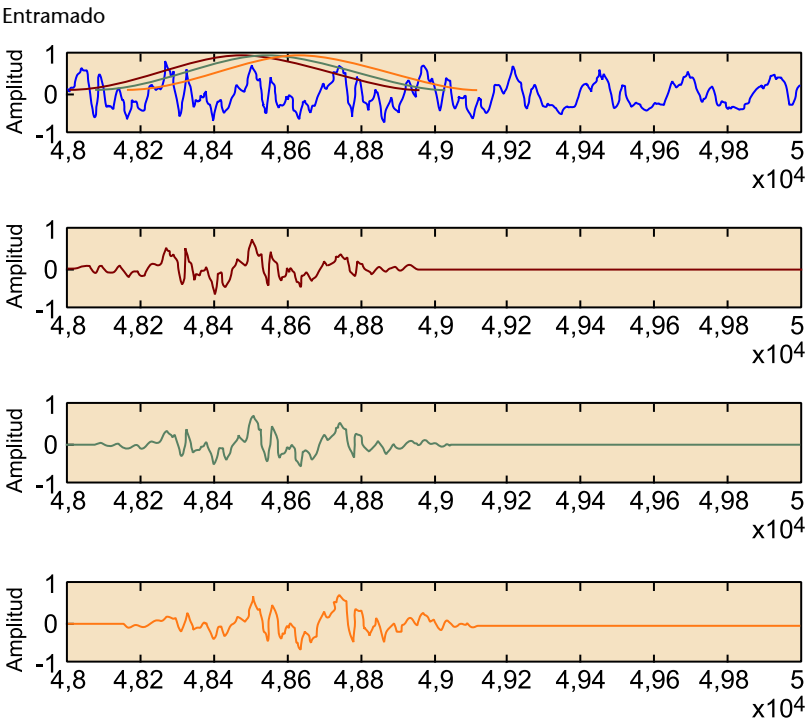

Figura 3. Procedimiento para el entramado de una señal de voz. La representación superior se corresponde con la señal de entrada, donde se han superpuesto las diferentes ventanas que se aplican durante el entramado. Las representaciones inferiores se corresponden con tres tramos de la señal superior.

La longitud de la ventana es el intervalo temporal en el que la voz se considera estacionaria, es decir, que las características de la voz no varían. Este intervalo de tiempo en el que las características de la voz son estables es normalmente de unos 25 milisegundos. Por lo tanto, la duración de la ventana se suele fijar en 25 milisegundos. Las ventanas más utilizadas para el entramado son la ventana de Hamming o la ventana de Hanning.

Ved también

Podéis consultar el módulo "Diseño y análisis de filtros en procesamiento de audio" para conocer más detalles sobre estas ventanas.

La distancia entre dos instantes de análisis consecutivos (es decir, la distancia entre los puntos centrales de las ventanas) es un valor fijo, habitualmente entre 10 y 12 milisegundos. Por lo tanto, estas ventanas se encabalgan. Hay que señalar que unos valores más grandes provocarían una pérdida de información, dado que no se podría seguir la evolución temporal de la señal de voz. Por el contrario, unos valores más pequeños aumentarían el número de tramos de análisis y, por lo tanto, de vectores de características, sin aportar información nueva para el reconocimiento.

En resumen, vemos que para tener un buen análisis y obtener unos vectores de características correctas se utilizan ventanas de 25 milisegundos que están encabalgadas en el tiempo. En la figura 3 se observa este encabalgamiento.

4.2. Estimación paramétrica

La **estimación paramétrica** consiste en calcular, para cada tramo, las características que se utilizarán en el reconocimiento. En la mayoría de los ASR se utilizan características espectrales (del dominio de la frecuencia), dado que son más fuertes ante el ruido que la misma señal en el dominio temporal y tienen una representación más compacta (menos coeficientes).

Robustez

En procesamiento del habla se utiliza el término **robusto** cuando se quiere indicar que un valor, el resultado de una técnica, etc., queda poco afectado por la presencia de ruido o por discrepancias entre las condiciones de entrenamiento y test.

Una de las características espectrales más utilizadas en el reconocimiento son los coeficientes *Mel cepstrum*, denominados también *mel-frequency cepstrum coefficients* (MFCC), junto a información sobre la variación que tienen en el tiempo (los parámetros delta) y el incremento de esta variación (los parámetros doble delta).

El cepstrum $\widehat{x}[n]$ de la secuencia temporal $x[n]$ se define como la transformada inversa de Fourier (IDFT) del logaritmo del valor absoluto del espectro de una señal (DFT de la señal):

$$\widehat{x}[n] = \mathrm{IDFT}(\log(|\mathrm{DFT}(x[n])|))$$

Cálculo de coeficientes cepstrum

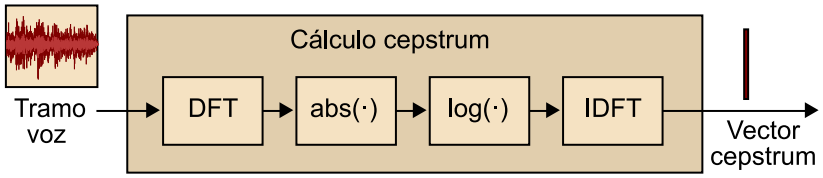

Figura 4. Procedimiento para calcular el vector de coeficientes cepstrum para un tramo de señal

A pesar de que los cepstrums son útiles, en procesamiento del habla normalmente se da un paso más y se transforma el espectro de la señal a la escala de Mel antes de calcular los cepstrum (es decir, antes de aplicar el logaritmo y el IDFT) para obtener los parámetros Mel cepstrum.

La **escala de Mel** es una escala perceptiva que relaciona hercios e índice de Mel, de tal manera que los índices de Mel representan tonos que perceptivamente son equidistantes[2]. Por lo tanto, la escala de Mel está construida siguiendo un esquema de funcionamiento parecido al oído humano. En la figura 5 encontramos una representación de la escala de Mel:

[(2)]Es decir, que están a la misma distancia.

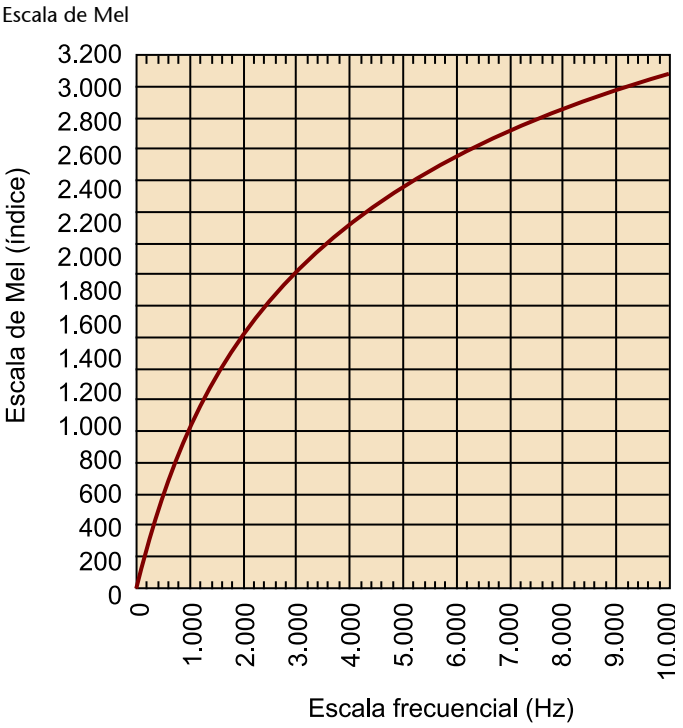

Figura 5. Representación de la relación entre hercios e índice de la escala de Mel

Así, el vector de características basado en MFCC para cada tramo tiene la estructura siguiente:

$$[\widehat{x}_0,\ \widehat{x}_1,\ldots,\ \widehat{x}_n,\ \Delta\widehat{x}_0,\ \Delta\widehat{x}_1,\ \ldots,\Delta\widehat{x}_n,\ \Delta^2\widehat{x}_0,\ \Delta^2\widehat{x}_1,\ \ldots,\Delta^2\widehat{x}_n]$$

Lectura de la fórmula

- $\widehat{x}_i$ son los coeficientes MFCC.
- $\Delta\widehat{x}_i$ son los parámetros delta.
- $\Delta^2\widehat{x}_i$ son los parámetros doble delta.

Los parámetros delta $\Delta\widehat{x}_i$ se calculan haciendo la primera derivada de los MFCC y los parámetros doble delta $\Delta^2\widehat{x}_i$ haciendo la segunda derivada.

El número habitual de coeficientes MFCC utilizados en los vectores de características se sitúa entre 10 y 15 para voz muestreada a 8 kHz (voz de calidad telefónica), dado que se considera que contienen la información más relevante para discriminar diferentes sonidos entre sí.

El estándar ETSI ES 202 050, de reconocimiento distribuido, denominado también *Aurora*, define los valores siguientes para la extracción del vector de características:

- Longitud de los tramos de voz: 25 ms.
- Distancia entre los instantes de análisis: 10 ms.
- Dimensión del vector de características: 39 componentes (13 coeficientes MFCC, 13 deltas y 13 dobles deltas) para voz muestreada a 8 kHz.

5. El módulo de descodificación

El objetivo del módulo de descodificación es encontrar la secuencia de palabras que con más probabilidad ha generado la secuencia de vectores de características acústicas extraídas en el módulo anterior.

La frase anterior se puede escribir en lenguaje matemático de la manera siguiente. Si definimos la secuencia de características acústicas de entrada como $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_N$, donde el subíndice N es el número total de ventanas utilizadas, y una secuencia de palabras como $W = \boldsymbol{w}_1\boldsymbol{w}_2 \ldots \boldsymbol{w}_m$, el módulo de descodificación busca la secuencia de palabras $\widehat{W}$ que maximice la probabilidad condicional dadas las características acústicas de entrada:

$$\widehat{W} = arg\max_{w} P(W|\mathrm{X})$$

La búsqueda de $\widehat{W}$ no se puede llevar a cabo directamente con la ecuación anterior por la complejidad que tiene, pero se puede simplificar mediante unas cuantas operaciones matemáticas y los conocimientos que tenemos a priori del lenguaje y el habla. Aplicando el teorema de Bayes, la probabilidad de una secuencia de palabras dada la secuencia acústica se puede escribir dependiendo de la probabilidad de la secuencia acústica dada una secuencia de palabras:

$$\widehat{W} = arg\max_{w} P(W|X) = arg\max_{w} \frac{P(W)\mathrm{P}(X|W)}{P(X)}$$

Teorema de Bayes

Si $\{A_1, A_2, \ldots, A_n\}$ es un conjunto de sucesos mutuamente excluyentes y exhaustivos, de modo que la probabilidad de cada uno es diferente de 0, y si B es un suceso cualquiera del que se conocen las probabilidades condicionales $P(B \mid A_i)$, entonces la probabilidad $P(A_i|B)$ viene dada por la expresión siguiente:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde $P(A_i)$ es la probabilidad de que suceda A_i, $P(B|A_i)$ es la probabilidad de B, sabiendo que ha sucedido A_i y $P(A_i|B)$ es la probabilidad a posteriori de A_i, o, dicho de otro modo, la probabilidad de que suceda A_i sabiendo que ha sucedido B.

La probabilidad de la secuencia acústica de entrada $P(\boldsymbol{X})$ no depende de $\widehat{W}$, dado que viene dada por la señal acústica de entrada en el ASR. Por lo tanto, el término $P(\boldsymbol{X})$ se puede eliminar de la ecuación anterior. Así se obtiene la expresión final del módulo de descodificación:

$$\widehat{W} = arg\max_{w} P(X|W)P(W)$$

En la ecuación final el espacio de búsqueda ha quedado definido por dos términos:

- $P(X|W)$: la probabilidad de que se genere la secuencia de características acústicas $\boldsymbol{X}$ proporcionada por el módulo anterior dada una secuencia de palabras W. Este término se denomina *modelo acústico*.

- $P(W)$: la probabilidad de la secuencia de palabras W. Este término se denomina *modelo de lenguaje*.

Así, el descodificador se puede representar con el diagrama de bloques siguiente:

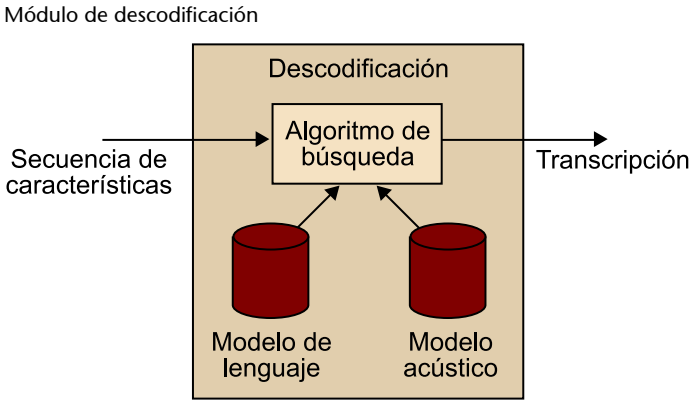

Figura 6. Diagrama de bloques del módulo de descodificación

Observad que el descodificador está formado por dos tipos de elementos: un bloque de cálculo (algoritmo de búsqueda) y modelos probabilísticos (modelo de lenguaje y modelo acústico).

El bloque de cálculo está activo cuando se ejecuta un reconocimiento, utilizando el modelo de lenguaje y el modelo acústico para obtener la información necesaria para hacer la búsqueda.

Los modelos probabilísticos no se construyen durante la ejecución del reconocimiento, sino que se han construido previamente en la fase de entrenamiento del ASR.

En los apartados siguientes describiremos qué son y cómo se construyen los modelos acústicos durante la fase de entrenamiento (apartado 5.1, "Modelado acústico"); qué son y cómo se construyen los modelos de lenguaje durante la fase de entrenamiento (apartado 5.2, "Diccionarios y modelos de lenguaje"), y finalmente cómo se resuelve la ecuación final en la fase de ejecución, es decir, cómo trabaja el bloque de cálculo (o algoritmo de búsqueda) para encontrar la mejor transcripción posible para la secuencia de características de entrada al módulo de descodificación (apartado 5.3, "Algoritmos de búsqueda").

5.1. Modelado acústico

El objetivo del modelado acústico es crear un conjunto de modelos probabilísticos que representen todos los sonidos del lenguaje que se deben reconocer.

Para crear los modelos acústicos se ha de determinar:

- Qué unidades acústicas (qué sonidos) se quieren representar.
- Qué modelo probabilístico se utilizará.
- El método de estimación de los modelos.

5.1.1. Unidades acústicas

Las unidades más utilizadas para el modelado acústico son los **trifonemas**. Los trifonemas son fonemas dependientes del contexto, es decir, son los fonemas pero considerando unidades diferentes a los fonemas que no tienen igual el fonema anterior y posterior.

Ved también

Para más información, podéis consultar el módulo "Introducción al habla".

Los trifonemas son unidades adecuadas para realizar reconocimiento del habla puesto que no solamente representan las características estables de la realización central de los fonemas, sino que también capturan los efectos coarticulatorios anteriores y posteriores provocados por el contexto en el que se encuentran.

5.1.2. Modelo probabilístico

El modelo más utilizado para representar las unidades acústicas para el reconocimiento son los modelos ocultos de Markov o *hidden Markov models* (HMM).

Los HMM modelan secuencias de vectores y representan no solo cada vector individualmente, sino también la evolución de la secuencia. Por lo tanto, son muy adecuados para describir situaciones que evolucionan en el tiempo, como es el caso de la voz.

Los HMM están formados por tres elementos diferentes: estados, transiciones y símbolos emitidos. En la figura 7 podéis ver la representación de un HMM de un trifonema, donde los estados s_i se han representado por medio de nodos, las transiciones por medio de enlaces entre nodos y los símbolos emitidos se representan con la letra $\boldsymbol{x}$.

Modelo oculto de Markov de tres estados

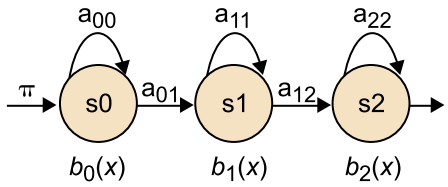

Figura 7. Modelo oculto de Markov de tres estados

Para cada incremento de tiempo, el modelo lleva a cabo una transición entre estados y emite un símbolo nuevo. El comportamiento de un HMM está regido por dos tipos de funciones de probabilidad: la probabilidad de transición entre estados y la probabilidad de emisión de los símbolos.

La probabilidad de transición entre estados, representada como a_{ij} en la figura 7, es la probabilidad de que el modelo esté en estado *j* si en el instante anterior estaba en el estado *i*.

$$a_{ij} = P(s(t) = j | s(t-1) = i)$$

Ejemplo

Un ejemplo de las probabilidades de pasar del estado *i* al estado *j* del HMM de la figura 7, recogidas en una matriz de transición, es el siguiente:

$$A = \{a_{ij}\} = \begin{bmatrix} 0{,}2 & 0{,}8 & 0 \\ 0 & 0{,}15 & 0{,}85 \\ 0 & 0 & 0{,}3 \end{bmatrix}$$

Observad que la suma de los valores de cada fila de la matriz es igual a 1, dado que todas las posibles opciones de transición consisten en quedarse en el mismo estado o pasar al estado siguiente.

La probabilidad que falta en la última fila (de valor 0,7) representa la probabilidad de salir del modelo, es decir, la probabilidad de que el vector de características sea el último del trifonema representado por el HMM. De manera similar, también se define la probabilidad π de entrada al modelo, es decir, la probabilidad de que observemos el primer vector de características del trifonema.

La probabilidad de emisión de los símbolos, representada como $b_{s(t)}(x)$ en la figura 7, depende del estado en el que se encuentre el HMM. Las funciones de emisión más habituales son modelos de mezclas de gaussianas (GMM), denominadas también *gaussian mixture model*, por las buenas propiedades de representación que tienen de cualquier tipo de distribución.

GMM

Un GMM es una función de densidad de probabilidad construida como una suma ponderada de gaussianas:

$$p(x;\theta) = \sum_{q=0}^{Q-1} \alpha_q N(x;\theta_q) \quad \text{de manera que} \quad \sum_{q=0}^{Q-1} \alpha_q = 1$$

$N(x;\theta)$ representa una función gaussiana cuyos parámetros son $\theta = \{\mu, \sigma\}$, es decir, la esperanza μ y la desviación estándar σ:

$$N(x;\theta)=\frac{1}{(2\pi)^{p/2}}\left|\sigma\right|^{1/2}e^{\left[-\frac{1}{2}(x-\mu)^T\sigma(x-\mu)\right]}$$

Función gaussiana

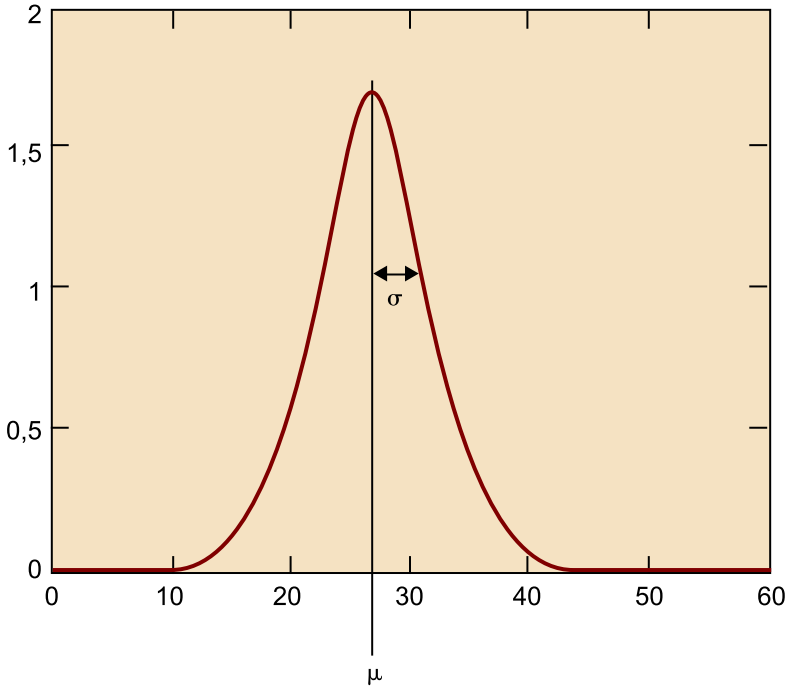

Figura 8. Representación de una función gaussiana

donde p es la longitud del vector de características $\boldsymbol{x}$.

Estos tipos de modelos se denominan *ocultos* porque podemos saber qué símbolo (es decir, qué vector de características MFCC) se ha emitido, pero no podemos saber en qué estado del modelo se encuentra.

Para el modelado de trifonemas se utilizan HMM de tres estados, para capturar las tres fases diferenciadas de los fonemas: coarticulación con el fonema anterior (estado s_0), la parte estable del fonema (estado s_1) y la coarticulación con el fonema posterior (estado s_2). Dado que las tres fases están presentes en todas las realizaciones, las transiciones permitidas solo son entre los estados 0 y 1, 1 y 2, además de cada estado consigo mismo. Los símbolos emitidos son los vectores de características MFCC.

5.1.3. Estimación del modelo acústico

La construcción del modelo acústico de un ASR significa valorar los parámetros de un HMM para cada trifonema de la lengua. Por lo tanto, tendremos tantos HMM de tres estados como trifonemas tenga la lengua en la que queremos reconocerlos.

Los parámetros que se deben estimar para cada HMM son los siguientes:

- La probabilidad de entrada al modelo: π.

- La matriz de las probabilidades de transición: $A = \{a_{ij}\}$ para $i = 0,1,2$ y $j = 0,1,2$.

- La media y la desviación estándar de cada una de las gaussianas de los GMM de los tres estados: $\theta_{qi} = \left\{\mu_{qi}, \sigma_{qi}\right\}$ para $i = 0,1,2$ y $q = 0... Q1$, donde Q es el número de gaussianas de los GMM.

Para estimar estos parámetros se utilizan datos reales de diferentes realizaciones de cada trifonema. Esto significa obtener grabaciones de múltiples personas, segmentar la voz en trifonemas y utilizar los vectores de características de todas las apariciones de un mismo trifonema para entrenar al HMM correspondiente.

La obtención de una base de datos (o corpus, en el argot de las tecnologías del habla) adecuada para el modelado acústico es una de las tareas más complejas de la construcción de un ASR y es la tarea que requiere más tiempo. Las especificaciones del corpus (duración total, número de locutores, distribución de sexo y edad de los locutores, canal de grabación, etc.) tienen una gran influencia en las prestaciones finales del ASR, dado que cuanto más parecidas sean las características acústicas de los modelos al audio de la aplicación final, mejores serán los resultados.

Corpus

En procesamiento del habla se denomina *corpus* a las bases de datos utilizadas para la construcción de modelos acústicos, de lenguaje o sistemas de síntesis del habla. El plural de *corpus* en español es invariable: *corpus*.

El diagrama de bloques por el proceso de estimación del modelo acústico se muestra en la figura 9.

Estimación del modelo acústico

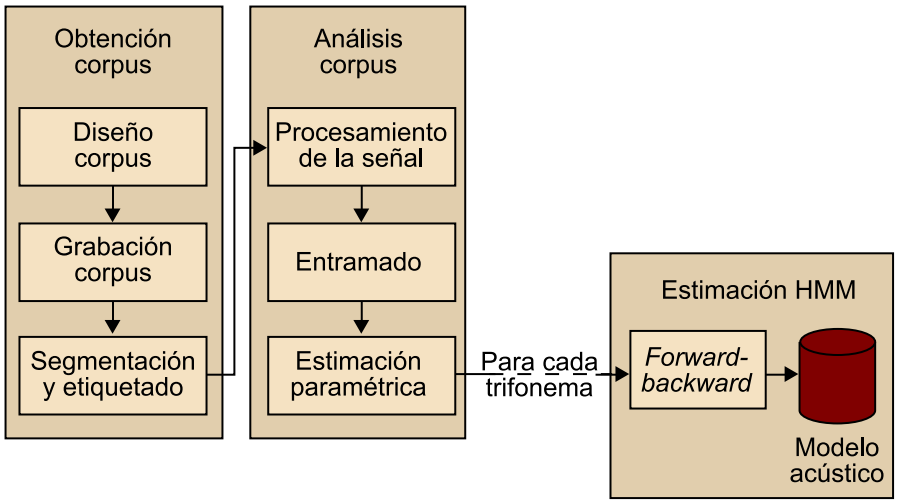

Figura 9. Diagrama de bloques del proceso de estimación del modelo acústico de un ASR en la fase de entrenamiento

El algoritmo que se utiliza para entrenar un HMM a partir de datos del corpus es el algoritmo iterativo de Baum-Welch, denominado también *forward-backward*.

Bibliografía complementaria

Para ampliar la información sobre la estimación de los modelos acústicos y el algoritmo de Baum-Welch, podéis consultar la referencia siguiente:

L. Rabiner; B. H. Juang (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall PTR.

El objetivo del modelado acústico es crear un conjunto de modelos probabilísticos que representen todos los sonidos del lenguaje que se han de reconocer.

Las unidades acústicas más utilizadas para representar los sonidos son los trifonemas, dado que tienen en cuenta los efectos coarticulatorios del contexto anterior, del contexto posterior y la parte central estable del fonema.

Los modelos probabilísticos más utilizados para el modelado acústico son los HMM, que modelan secuencias de vectores de características. Los vectores de características más utilizados son los MFCC.

Para construir un modelo acústico se debe estimar un HMM para cada trifonema de la lengua. Esta estimación se efectúa a partir de vectores de características MFCC reales, provenientes del análisis de un corpus de grabaciones. El algoritmo más utilizado para crear los HMM del modelo acústico a partir del corpus de entrenamiento es el de Baum-Welch.

5.2. Diccionarios y modelos de lenguaje

El objetivo de los modelos de lenguaje, junto a los diccionarios, es capturar las propiedades del lenguaje, concretamente qué palabras hay y cuál es la relación que tienen entre sí.

El **diccionario** es una lista que contiene todas las palabras que el ASR es capaz de transcribir, junto a la transcripción fonética de estas palabras. Cualquier palabra que no aparezca en la lista se denomina *palabra fuera del vocabulario* o *out of vocabulary word* (OOV *word*). El ejemplo más común de palabras OOV son los nombres propios.

Ejemplo de un diccionario de un servicio de atención telefónica de averías informáticas:

Palabra	Transcripción
técnico	[te'kniko]
cuenta	[kwe'Nta]
...	...

La medida del diccionario es un valor de compromiso. Un número elevado de palabras puede provocar más probabilidad de error, dado que el abanico de posibilidades de transcripción es más grande, pero un número muy reducido de palabras aumenta la probabilidad de que los usuarios utilicen palabras de fuera del diccionario.

Los **modelos de lenguaje** contienen información sobre el modo como se relacionan las palabras del diccionario, es decir, especifican todas las combinaciones posibles (con sentido semántico) que pueden crear las palabras del diccionario, junto con la probabilidad de cada combinación.

Los modelos estocásticos de lenguaje consisten en valorar la probabilidad $P(W)$ de una secuencia $W = w_1 w_2 ... w_m$ de palabras como:

$$P(W) = P(w_1, w_2, w_3 \dots w_m) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_m|w_1, w_2, w_3 \dots w_{m-1}) =$$

$$= \prod_{i=1}^{m} P(w_i|w_1, w_2, w_3 \dots \ w_{i-1})$$

Por tanto, $P(W)$ nos informa de la frecuencia de aparición de la frase W en el lenguaje que analizamos.

Los *n*-gramas simplifican la ecuación anterior, reduciendo la memoria del modelo, de manera que la aparición de una palabra solo depende de las *n* palabras anteriores. Los *n*-gramas más utilizados en ASR son los bigramas ($n = 2$) y los trigramas ($n = 3$).

Ejemplo

En un modelo de lenguaje formado por bigramas, la probabilidad de la frase "este es el tema ocho" se calcula como sigue:

$$P(este \ \ es \ \ el \ \ tema \ \ ocho) = P(este|< null >)P(es|este)P(el|es)P(tema|el)P(ocho|tema)P(< null > |ocho)$$

donde <null> representa comienzo o final de frase.

Como en el caso de los modelos acústicos, los *n*-gramas se entrenan a partir de textos reales del ámbito de la aplicación para la que se utilizará el ASR. La probabilidad de cada *n*-grama se calcula de la manera siguiente:

$$P(w_m|w_{m-n} \dots w_{m-1}) = \frac{C(w_{m-n} \dots w_{m-1}, w_m)}{C(w_{m-n} \dots w_{m-1})}$$

donde $C(W)$ indica el número de veces que ha aparecido la secuencia de W palabras en el texto de entrenamiento de los *n*-gramas.

Ejemplo

Si el corpus de entrenamiento del modelo de lenguaje es el siguiente:

[*este es el tema ocho*;

este tema es el tema ocho]

los bigramas que se pueden entrenar son estos:

***P*(*este*|<null>)**	2 / 2 = 1
***P*(*el*|*es*)**	2 / 2 = 1
***P*(*es*|*este*)**	1 / 2 = 0,5
***P*(*es*|*tema*)**	1 / 3 = 0,33
***P*(*tema*|*el*)**	2 / 2 = 1
***P*(*tema*|*este*)**	1 / 2 = 0,5
***P*(*ocho*|*tema*)**	2 / 2 = 1
***P*(<null>|*ocho*)**	2 / 2 = 1

Para entrenar *n*-gramas se necesita un texto de entrenamiento de unos cuantos millones de palabras, y aun así no pueden capturar todas las combinaciones posibles. Por ello se han diseñado técnicas de suavización. Estas técnicas reservan probabilidad para combinaciones de palabras no vistas en el entrenamiento, para asignarles una probabilidad diferente de 0 y así reconocerlas.

Bibliografía complementaria

Para saber más sobre esto, podéis consultar las referencias siguientes:

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing* (pág. 562-573). Englewood Cliffs, NJ: Prentice Hall PTR.

Juegos de herramientas de código abierto para crear modelos de lenguaje:

IRSTLM

SRILM

5.3. Algoritmos de búsqueda

Una vez hemos explicado qué son los modelos acústicos y los modelos de lenguaje y cómo se construyen en la fase de entrenamiento de un ASR, estamos preparados para ver cómo se debe hacer la descodificación de la secuencia de vectores de características y obtener la transcripción del audio a partir de la siguiente ecuación. Recordémosla:

$$\widehat{W} = arg\max_{W} P(X|W)P(W)$$

Para resolver esta ecuación, se debe buscar cuál es la secuencia de palabras $\widehat{W}$ que, entre todas las posibles secuencias W del modelo de lenguaje, proporciona una probabilidad mayor cuando tenemos la realización acústica X. Por lo tanto, el algoritmo que se debe utilizar para resolver la ecuación es un algoritmo de búsqueda.

Para llevar a cabo la búsqueda, antes de nada se ha de construir el espacio de búsqueda, es decir, expandir el modelo de lenguaje para tener todas las posibles transcripciones que se pueden reconocer con la ASR.

Si tenemos una aplicación con un diccionario que contiene tres palabras {*la*, *casa*, *roja*} y el modelo de lenguaje es un bigrama definido por las probabilidades siguientes:

***P*(·,·)**	***la***	***casa***	***roja***	**Final_frase**
Inicio_frase	0,8	0	0,2	0
la	0	0,7	0,3	0
casa	0	0	0,75	0,25
roja	0	0	0	1

el espacio de búsqueda será el siguiente:

Espacio de búsqueda

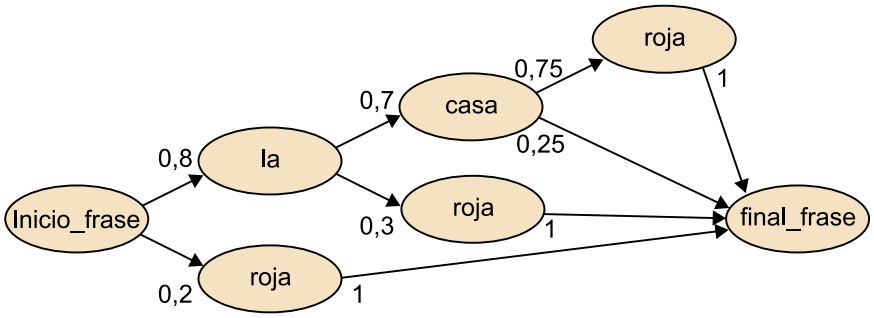

Figura 10. Espacio de búsqueda de un modelo de lenguaje

Una vez construido el espacio de búsqueda se calcula la probabilidad de cada rama, es decir, la probabilidad de cada una de las posibles transcripciones según el modelo de lenguaje. Así, obtenemos el término $P(X|W)$ de la ecuación anterior.

Siguiendo el ejemplo anterior, se calculan las probabilidades:

- P("<Inicio_frase> la casa roja <Final_frase>") = P(la|(<Inicio_frase>) P(casa|la) P(roja|casa) P(<Final_frase>|roja) = 0,8 × 0,7 × 0,75 × 1 = 0,42
- *P*(("<Inicio_frase> *la casa* <Final_frase>") = *P*(*la* |(<Inicio_frase>) *P*(*casa*| *la)P*(<Final_frase>| *casa*) = 0,8 × 0,7 × 0,25 = 0,14
- *P*(("<Inicio_frase> *la roja* <Final_frase>") = *P*(*la* |(<Inicio_frase>) *P*(*roja*| *la)P*(<Final_frase>| *roja*) = 0,8 × 0,3 × 1 = 0,24
- *P*(("<Inicio_frase> *roja* <Final_frase>") = *P*(*roja*|(<Inicio_frase>) *P*(<Final_frase>| *roja*) = 0,2 × 1 = 0,2

El resto de las combinaciones tienen una probabilidad cero.

Para incorporar la información acústica, primero se sintetiza el modelo acústico de cada rama del espacio de búsqueda concatenando los modelos de trifonemas necesarios para formar las frases, según la transcripción detallada en el diccionario y las reglas fonéticas de la lengua.

Los modelos acústicos sintetizados para cada rama de la figura 10 son los siguientes:

- Rama "la casa roja":
  [l+a l-a+k a-k+a k-a+s a-s+a s-a+ř a-ř+o ř-o+x o-x+a x-a]

- Rama “la casa”: [l+a l-a+k a-k+a k-a+s a-s+a s-a]
- Rama “la roja: [l+a a-ř+o ř-o+x o-x+a x-a]
- Rama “roja”: [ř+o ř-o+x o-x+a x-a]

Después, se evalúan los modelos acústicos sintetizados para calcular la probabilidad de la secuencia de vectores de características de entrada dada la transcripción de cada rama del espacio de búsqueda. Así se obtiene término $P(X|W)$ de la ecuación anterior.

¿Cómo se evalúa un modelo acústico?

Recordemos que los modelos acústicos habitualmente son conjuntos de HMM de tres estados (cada HMM es un trifonema). Por tanto, $P(X)$ del trifonema T se puede escribir de la siguiente forma:

$$P(X)=\sum_{r=1}^{R} P_T(X|s_r)P(s_r)\approx P_T(X|s_{r\max})P(s_{r\max})$$

donde s_r indica una posible secuencia de estados del HMM que generen X, y donde R es el número de secuencias posibles, y $r_{\max}$ es la secuencia de estados que proporciona la máxima probabilidad. $P_T(X|s_r)$ se calcula como sigue:

$$P_T(X|s_r)=\prod_{n=0}^{N-1} bs_{r(n)}(x)$$

donde $bs_{r(n)}(x)$ es la probabilidad de emisión del vector de características $\boldsymbol{x}$ y N la longitud de la secuencia de estado s_r.

Finalmente, la multiplicación de los $P(X)$ de los distintos HMM que configuran el modelo acústico de cada rama da el valor del modelo acústico de la rama, es decir, $P(X|W)$.

Finalmente se calcula la probabilidad de cada rama multiplicando la probabilidad obtenida con el modelo de lenguaje $P(W)$y la probabilidad obtenida mediante los modelos acústicos $P(X|W)$. La transcripción asociada a la rama con mayor probabilidad es la elegida como transcripción óptima.

Seguro que habéis observado que la búsqueda de la rama más probable requiere un conjunto de operaciones que implican un gran coste de cálculo, dado que el espacio de búsqueda puede ser enorme. Por esta razón, el procedimiento de búsqueda directa no se puede utilizar en ASR, donde el vocabulario es mediano o grande. En estos casos, en lugar de expandir todo el espacio de búsqueda se utilizan algoritmos basados en programación dinámica junto con técnicas de poda o *pruning* para reducir el número de ramas del espacio de búsqueda que se deben explorar. El algoritmo de Viterbi es el más utilizado en los ASR.

Poda

El término *poda* o *pruning* se utiliza para referirse a las técnicas que limitan el número de caminos activos en la búsqueda por Viterbi. Por lo tanto, gracias a esta técnica, no se exploran todos los caminos de un espacio de búsqueda, de modo que se ahorran muchos cálculos en espacios de búsqueda grandes. Sin embargo, se garantiza que se encontrará una solución similar a la que se encontraría si se exploraba todo el espacio de búsqueda.

Bibliografía complementaria

Para saber más sobre el tema, podéis consultar las referencias siguientes:

L. Rabiner; B. H. Juang (1993). *Fundamentals Of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall PTR.

Juego de herramientas de código abierto para trabajar con modelos HMM y el reconocimiento del habla:

HTK

El módulo de descodificación obtiene la secuencia de palabras que con más probabilidad ha generado la secuencia de características acústicas extraídas del audio.

La descodificación utiliza dos fuentes de informaciones diferentes para resolver una ecuación de búsqueda: los modelos acústicos y los modelos de lenguaje más diccionario.

El modelo acústico de un ASR habitualmente es un conjunto de HMM. Cada HMM representa la secuencia de vectores de características (por ejemplo, MFCC) de una unidad acústica (por ejemplo, un trifonema) dependiendo de los estados del modelo.

El diccionario contiene todas las palabras que puede transcribir el ASR, junto a la transcripción fonética que tienen.

Los modelos de lenguaje capturan las probabilidades de combinación de las palabras del diccionario. Los modelos más utilizados son los *n*-gramas.

La búsqueda se efectúa teniendo en cuenta todas las posibilidades de transcripción y eligiendo la más probable según el audio de entrada, el modelo acústico y el modelo de lenguaje por medio del algoritmo de Viterbi.

6. Técnicas de adaptación

Los datos de entrenamiento de los modelos acústico y de lenguaje deben representar adecuadamente el ámbito de aplicación del ASR para tener unas buenas prestaciones. Por ejemplo, en la construcción de un ASR para atención telefónica no se han de utilizar grabaciones hechas directamente con un micrófono, sino con un teléfono, dado que las características espectrales de las señales son muy diferentes y aumentaría la tasa de error de la transcripción.

Aun así, por muy bien que se elijan los datos de entrenamiento, en el reconocimiento del habla siempre aparecerán situaciones nuevas: cambios en el micrófono o teléfono, en el estado del locutor (constipados, por ejemplo), en el ruido de fondo, etc. Para hacer frente a estos desacoplamientos aparecieron las técnicas de adaptación.

Las técnicas de adaptación utilizan una parte de los datos de entrada (audio que se debe reconocer) o de salida (transcripción automática) para mejorar con el uso el modelo acústico o el modelo de lenguaje, y que así puedan representar mejor una situación nueva. Por lo tanto, cuanto más se utiliza un ASR que incorpore técnicas de adaptación, mejor es el comportamiento que tiene ante situaciones nuevas, dado que cada uso que se hace del ASR es un entrenamiento, es decir, una posibilidad para aprender y mejorar.

Las estrategias de adaptación se clasifican según el impacto que causan al usuario de los ASR en:

- Adaptación intrusiva: el usuario debe hacer alguna acción (repetir una frase en concreto, permanecer sin decir nada para capturar el ruido de fondo, etc.).

- Adaptación no intrusiva: el usuario no es consciente de que el sistema se adapta a él y a sus condiciones acústicas.

Según cómo utilizan la salida del ASR para realizar la adaptación, se clasifican dependiendo de si se ha revisado la transcripción de salida o no en:

- Adaptación supervisada: las transcripciones que se utilizan para la adaptación están revisadas manualmente, es decir, un experto ha corregido los errores de transcripción del ASR antes de realizar la adaptación.

- Adaptación no supervisada: las transcripciones utilizadas para la adaptación son las que da directamente el ASR, que pueden contener errores.

Las técnicas más utilizadas en cualquiera de las estrategias de adaptación en los ASR son las que adaptan los modelos acústicos, concretamente *maximum a posteriori* (MAP) y *maximum likelihood linear regression* (MLLR). Las dos técnicas modifican los parámetros de los HMM.

A pesar de que existen técnicas de adaptación de modelos de lenguaje, el uso que se hace de ellas no está tan extendido, dado que la reducción de la tasa de error que implica es más pequeña que la reducción asociada a las técnicas de adaptación acústica.

Bibliografía complementaria

Para ampliar información sobre las técnicas de adaptación, podéis consultar la referencia siguiente:

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing*. Englewood Cliffs, NJ: Prentice Hall PTR (pág. 444-452).

7. Evaluación de la transcripción automática

Para saber cuál es la calidad de un ASR, se debe transcribir de manera automática un conjunto de audios y se ha de evaluar la cantidad de errores que introduce el ASR en las transcripciones automáticas respecto a unas transcripciones de referencia realizadas de manera manual por personas. Cuanto más baja es la tasa de error, es decir, cuanto más se asemeja la transcripción automática a la transcripción manual, mejor es el ASR en ese dominio.

En la transcripción proporcionada por el ASR se encuentran tres tipos de errores:

- Sustituciones: cuando una palabra se transcribe por otra diferente.
- Omisiones: cuando desaparece una palabra en la transcripción automática.
- Inserciones: cuando la transcripción automática incluye una palabra que no aparece en el audio original y a la vez no sustituye ninguna.

En el ejemplo siguiente se muestra una transcripción de referencia (lo que se dice realmente en el audio) y una transcripción automática (proveniente de un ASR), en la que se observan los tres tipos de errores. Los asteriscos indican los errores de omisión, la negrita los errores de inserción y el subrayado los errores de sustitución.

Transcripción de referencia: a la reunión asistieron diez personas.

Transcripción automática: * la reunión **ya** asintieron diez personas.

Una de las medidas de cálculo de la tasa de error más utilizada para evaluar los ASR es la tasa de error por palabra o *word error rate* (WER):

$$\mathrm{WER} = 100\% \frac{\mathrm{Subs} + \mathrm{Dels} + \mathrm{Ins}}{\text{Núm. palabras de la transcripción de referencia}}$$

La figura 11 muestra la evolución de las tasas de error de las evaluaciones del NIST durante los últimos veinte años.

NIST

El National Institute of Standards and Technology (NIST) es una agencia del Departamento de Comercio de los Estados Unidos de América que organiza anualmente evaluaciones de tecnología abiertas a todas las instituciones públicas y privadas. Las evaluaciones del NIST sirven para realizar una comparativa de los diferentes sistemas y una evaluación de la evolución de la tecnología.

Evolución de la WER en las evaluaciones del NIST

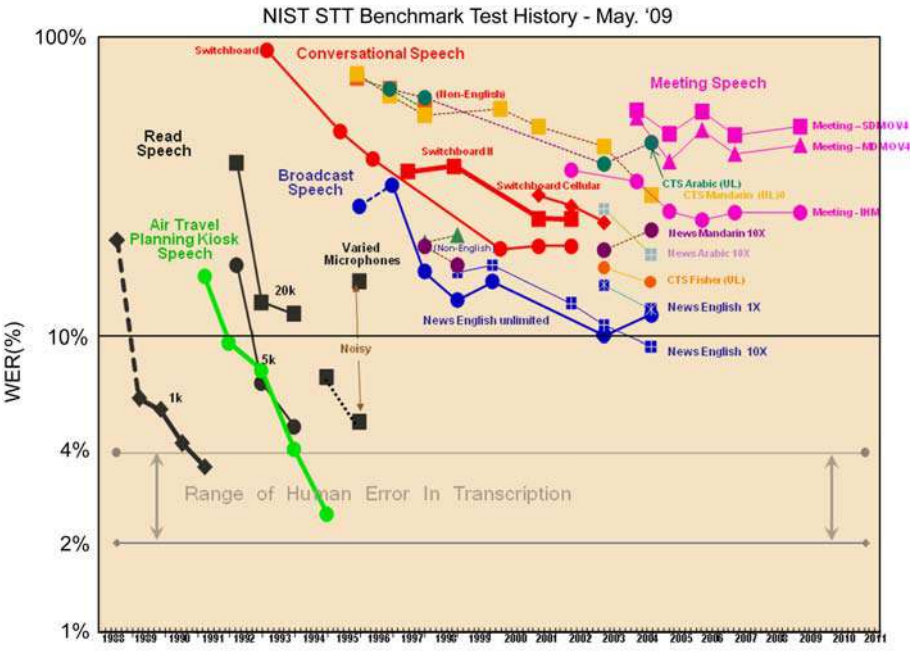

Figura 11. Evolución de la WER en las evaluaciones del NIST (figura extraída de la página web del NIST)

Actualmente, los ASR tienen prestaciones similares a los interlocutores humanos (WER por debajo del 5%) solo para ámbitos totalmente controlados, como el reconocimiento de una palabra entre una lista de opciones, siempre que el nivel de ruido presente en el audio sea bajo. En otros ámbitos favorables, como son los sistemas de dictado adaptados a un solo locutor o el reconocimiento de números de teléfono, los valores de la WER pueden estar en torno al 20%. Finalmente, para tareas más abiertas, tanto por el lenguaje utilizado como por el número de locutores que están implicados, como por ejemplo transcripciones de reuniones o de conversaciones telefónicas, los valores de la WER van desde el 50 hasta el 80%, dependiendo de la calidad del audio o del vocabulario utilizado.

La construcción de un ASR general, aplicable a cualquier locutor y a cualquier ámbito, es un reto todavía no resuelto.

Los ASR transforman la voz en texto.

Los ASR están formados por tres módulos: el módulo de procesamiento de la señal (encargado de limpiar el audio e identificar los segmentos en los que hay voz), el extractor de características (encargado de transformar el audio en una secuencia de parámetros MFCC) y el módulo de descodificación (encargado de hacer la búsqueda de la transcripción que maximiza la probabilidad de la secuencia de MFCC). La figura siguiente muestra un diagrama de bloques completo de un ASR:

Reconocedor automático del habla

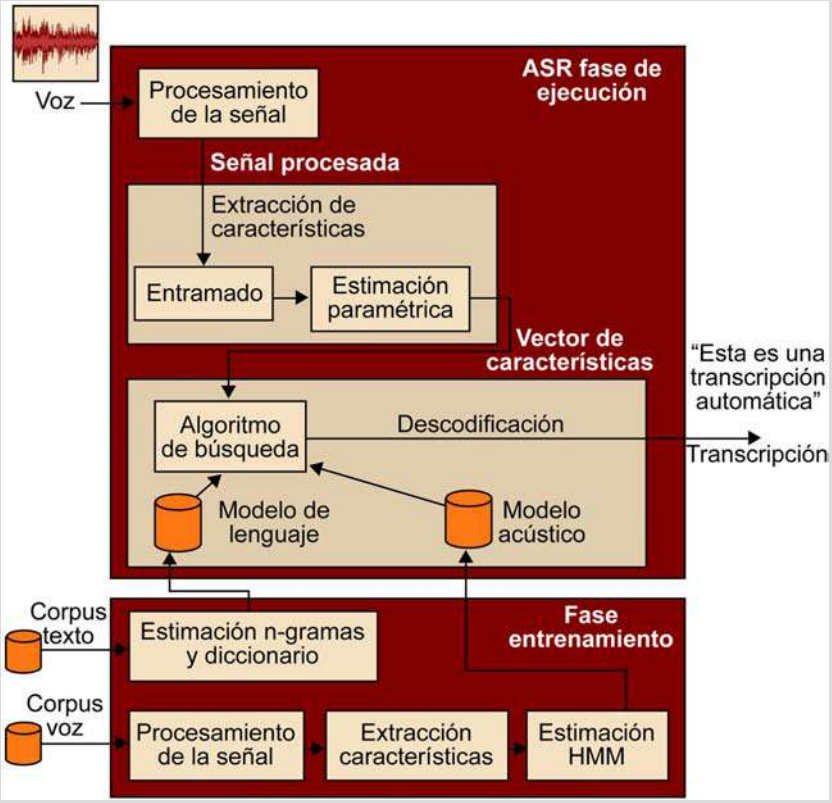

Figura 12. Diagrama de bloques completo de un ASR

La descodificación utiliza dos fuentes de informaciones diferentes para resolver la ecuación de búsqueda: los modelos acústicos y los modelos de lenguaje más el diccionario.

El modelo acústico de un ASR es un conjunto de HMM. Cada HMM representa la secuencia de vectores de características de un trifonema dependiendo de los estados del modelo.

El diccionario contiene todas las palabras que puede transcribir el ASR, junto a la transcripción fonética que tienen. Los modelos de lenguaje capturan las probabilidades de combinación de las palabras del diccionario. Los modelos más utilizados son los *n*-gramas.

Hay técnicas de adaptación que utilizan una parte de los datos de funcionamiento para ajustar el modelo acústico o el modelo de lenguaje a situaciones nuevas, con el fin de disminuir los errores en la transcripción.

A pesar de que los ASR tienen prestaciones aceptables para aplicaciones comerciales en dominios controlados, la construcción de un ASR universal es un reto todavía no resuelto.