Reconocimiento Automático del Habla / Tecnologías del Lenguaje Humano

2023-2024

Introducción

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

MIARFID
Máster en Inteligencia Artificia Reconocimiento de Formas e Imagen Digital

Comunicación hombre-máquina

- *Reconocimiento automático del habla* = Búsqueda de la secuencia de las palabras de una pronunciación. ***What did they say?*** *Buenos días, me llamo María José*
- *Comprensión del habla* = Búsqueda del significado de una pronunciación. ***What does it mean?*** *Saludo; información: nombre= "María José"*
- *Traducción del habla* = Conversión de una pronunciación en una secuencia de palabras en otro idioma. ***How do you say it in another language?*** *Good morning, my name is María José*
- Otros

  Text-To-Speech (Síntesis)

  Sistemas conversacionales

  Speaker recognition: ***Who did speak?***

  Speaker diarization: ***Who spoke when?***

  Paralinguistic aspects: ***How did they say it?*** (timing, intonation, voice quality)

  Sentiment análisis: ***How does the speaker feel?***

  Speech analytics

Qué es el Reconocimiento Automático del Habla

El Reconocimiento Automático del Habla (RAH) es la disciplina que se encarga de la concepción y realización de sistemas automáticos para la conversión automática de señales acústicas procedentes de un locutor humano en (secuencias de) categorías lingüísticas de un universo dado.

El RAH es un problema multidisciplinar, relacionado con

- procesamiento de la señal
- acústica
- psicología
- teoría de la comunicación y de la información
- lingüística
- fisiología
- informática (especialmente reconocimiento de formas e inteligencia artificial)

¿Por qué RAH? Ventajas y desventajas

Ventajas

- Forma natural de comunicación humana
- El habla es más rápida que la escritura
- Algunos canales (teléfono) son específicos para el habla
- Manos/Ojos libres para otras tareas
- Portabilidad (los micrófonos son más pequeños que los teclados)
- Funciona en ambientes oscuros
- Los locutores se pueden mover mientras hablan
- No se necesita práctica para utilizarlo

Desventajas

- ~~Tasas de reconocimiento aún bajas~~ No está completamente libre de errores
- Ambientes ruidosos/acústica en el agua
- ~~Aún más caros que los teclados~~
- No utilizable cuando se requiere silencio
- Privacidad de los datos registrados es motivo de preocupación

Tecnologías de habla en nuestro día a día

Asistentes personales

- Hacer llamadas, enviar mensajes, dictar correos
- Activar alarmas y reuniones
- Ayuda para la navegación (búsqueda de restaurantes…)
- Tomar notas
- Activar y/o identificar música
- Búsqueda por voz en navegadores

Tecnologías de habla en nuestro día a día

Dictado de mensajes (correos electrónicos, WhatsApp)

Tecnologías de habla en nuestro día a día

Servicio automático de atención al cliente

Tecnologías de habla en nuestro día a día

Proveedores

Microsoft

Google

amazon

NUANCE

VERINT

AVAYA

Hay muchas otras aplicaciones…

MEDIA - Subtitulado automático

CT
COMISIONADO DE TRANSPARENCIA
DE CASTILLA Y LEÓN

PORTAL DE TRANSPARENCIA
Y GOBIERNO ABIERTO
REGIÓN DE MURCIA

Portal de TRANSPARENCIA

Hay muchas otras aplicaciones…

MEDIA – Indexación automática

Hay muchas otras aplicaciones…

MEDIA – Transcripción automática

En España, hay
- 17 parlamentos
- 8.124 ayuntamientos

Mallol TRADUCTORS ASSOCIATS

serikat

séneca smart media

AMBISER

grupovitelsa

eCityclic

semic Effective IT Solutions

Actadigital

Hay muchas otras aplicaciones…

SEGURIDAD – Biometría de voz

ba
biometría aplicada

φFacePhi
Beyond Biometrics

Cestel

enigmedi

Powered by smowl ©

Hay muchas otras aplicaciones…

INDUSTRIA

Picking por voz

Dictado de informes de evaluación, error y mantenimiento
Apoyo operacional oral

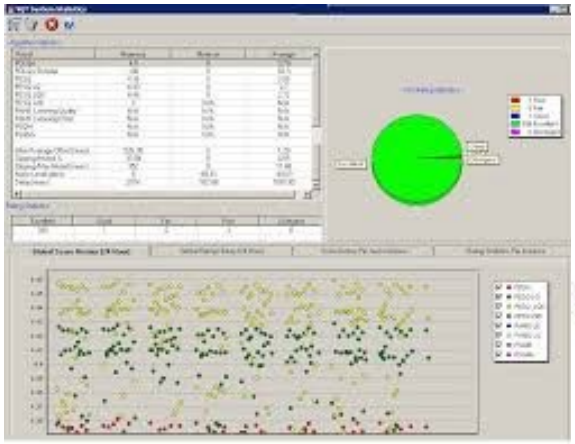
Hay muchas otras aplicaciones…

INDUSTRIA – Robots cooperativos

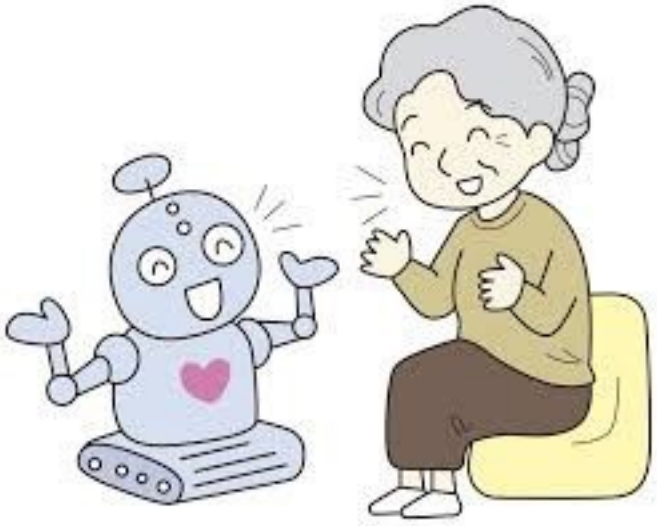
Hay muchas otras aplicaciones…

SALUD

Transcripción de informes médicos



Diagnosis y mejora de trastornos de la voz



Asistentes personales de salud

Deep Learning en RAH

Las Redes Neuronales Artificiales “irrumpieron” en el RAH a finales del siglo pasado…

- Clasificadores discriminativos no lineales en sistemas RAH a finales del siglo 20
- Menos restricciones en forma de características de entrada de la señal
- Avances en el hardware permiten nuevas aproximaciones (revolucionarias) en RAH


DEEP
LEARNING
CACHARRERÍA
El Planeta

El problema del RAH

Un problema difícil
Desde una perspectiva lingüística

El RAH es un problema complejo debido a:

- *Bidireccionalidad*: suele ser un proceso de diálogo.
- *Incompletitud*: se intercambia más información de la transmitida.
- *Continuidad*: Las marcas de separación de elementos (fonemas, sílabas, palabras, frases, etc.) que creemos percibir no existen.
- *Redundancia*: aunque se transmiten unos 50 bits por segundo de información, la señal requiere 100.000 bits por segundo.
- *Transitoriedad*: Hay mucha información en zonas transitorias (consonantes, transiciones entre vocales, etc.).
- *Ambigüedad:*
  - Homofónica: hojear/ojear, huso/uso, al abad/alabad
  - Semántica: “Se puede ver Teruel volando hacia Madrid”.
  - Pragmática: “Time flies like an arrow”.

- *Variabilidad*: la voz es un fenómeno complejo y afectado por numerosas fuentes de variabilidad.
  - El **locutor**:
    - Diferentes locutores presentan diferencias fisiológicas y sociolingüísticas que hacen sus pronunciaciones muy diferentes (acentos diferentes, dialectos...)
    - El estado físico y emocional del locutor afectan a su voz.
  - Las "unidades elementales" (**fonemas**) son muy dependientes del contexto.
  - El **entorno** (ruidoso o no), posición y características del micrófono también son fuente de variabilidad acústica.
  - El **estilo**: ¿habla continua o aislada? ¿discurso o conversación espontánea?
  - **Vocabulario**: comandos, lenaguaje científico, expresiones coloquiales...
  - El **idioma**: hay una estimación de 7000 idiomas, la mayoría con recursos de entrenamiento insuficientes, cambio de idioma...

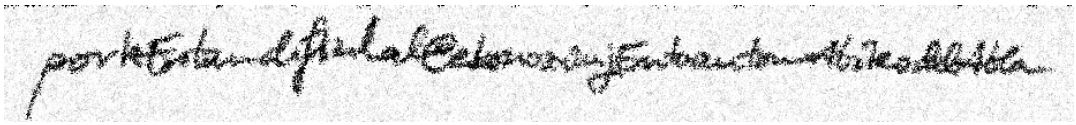
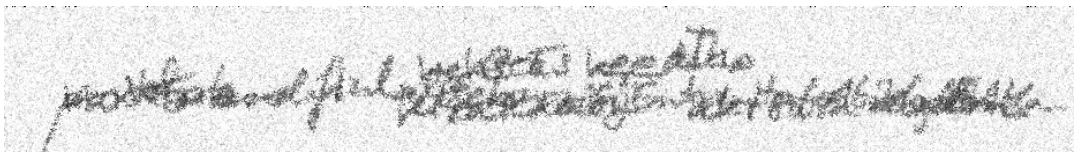
El problema del RAH

Un problema difícil
Desde una perspectiva *machine learning*

- Como problema de *clasificación:* de muy alta dimensionalidad.
- Como un problema *sequence-to-sequence* (entre secuencia acústica y secuencia de palabras): una entrada muy extensa.
- Los datos a menudo son ruidosos, con muchos factores "molestos" de variación en los datos.
- Cantidades muy limitadas de datos de entrenamiento disponibles (en términos de palabras) en comparación con la PLN basada en texto.
- La transcripción manual de voz es muy costosa.
- La naturaleza jerárquica y compositiva de la producción y comprensión del habla dificulta su manejo con un solo modelo.

El problema del RAH

Una analogía

Texto escrito	¿Por qué es tan difícil el Reconocimiento Automático del Habla?
Habla continua	Porquéestandifícilelreconocimientoautomáticodelhabla
Pronunciación	porkEstandifIzilel@ekonozimjEntoautomAtikodelAbla
Variabilidad acústica	
Ruido	
Efecto “fiesta cocktail”	

El problema del RAH

Un problema sencillo

¿Y por qué nos resulta tan sencillo a los humanos?

- Tenemos una gran capacidad de abstracción: una “a” es percibida como una “a” en cualquier contexto, aun cuando presenta muy diferentes realizaciones.
- Somos muy buenos segmentando sobre la marcha. Encontramos fonemas, sílabas, palabras, frases, etc...
- Disponemos de una gran base de datos de conocimiento internalizada sobre fonética, morfología, sintaxis, semántica, conocimiento pragmático, etc.
- Percibimos más de los que oímos: escuchamos fonemas que no han sido pronunciados, corregimos al vuelo errores sintácticos en lo dicho por el hablante, podemos entender completamente frases incompletas, etc...
- Somos fruto de la evolución. Hablar y entender el habla son procesos que han interaccionado entre sí en el propio proceso evolutivo y en el de construcción/invención del lenguaje hablado. Disponemos de un “órgano” del habla (Chomsky). Todos aprendemos a hablar a la misma edad, muy rápidamente, con poca información, cometiendo fallos similares. . .