CEU | Universidad San Pablo

UAM Universidad Autónoma de Madrid

audias

Audio, Data Intelligence and Speech

Search-on-Speech: Recuperación de información en repositorios de audio

Javier Tejedor Noguerales
Universidad San Pablo CEU
http://www.uspceu.com

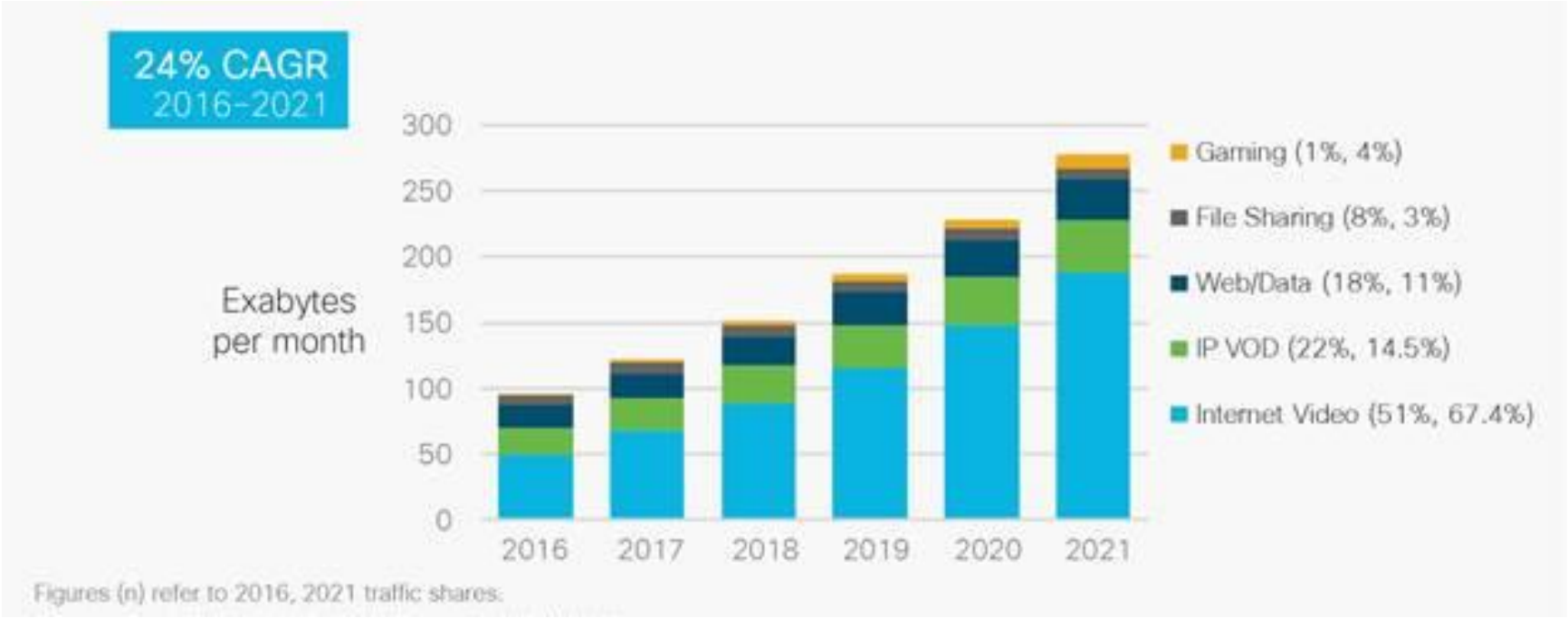
Doroteo Torre Toledano
Audias Research Group
Universidad Autónoma de Madrid
http://audias.ii.uam.es

Índice

- Motivación
- Concepto y tipos de búsqueda en voz
- Spoken Term Detection (STD)
- Query-by-Example STD (QbE STD)
- Evaluación
- Evaluaciones competitivas
- Evaluaciones ALBAYZIN Search-on-Speech 2012-18
- Conclusiones

Motivación

- Crecimiento del contenido audiovisual en Internet



Fuente: CISCO VNI Global IP Traffic Forecast 2016-21

- En 2021 el 82% del tráfico de Internet será audiovisual

Motivación

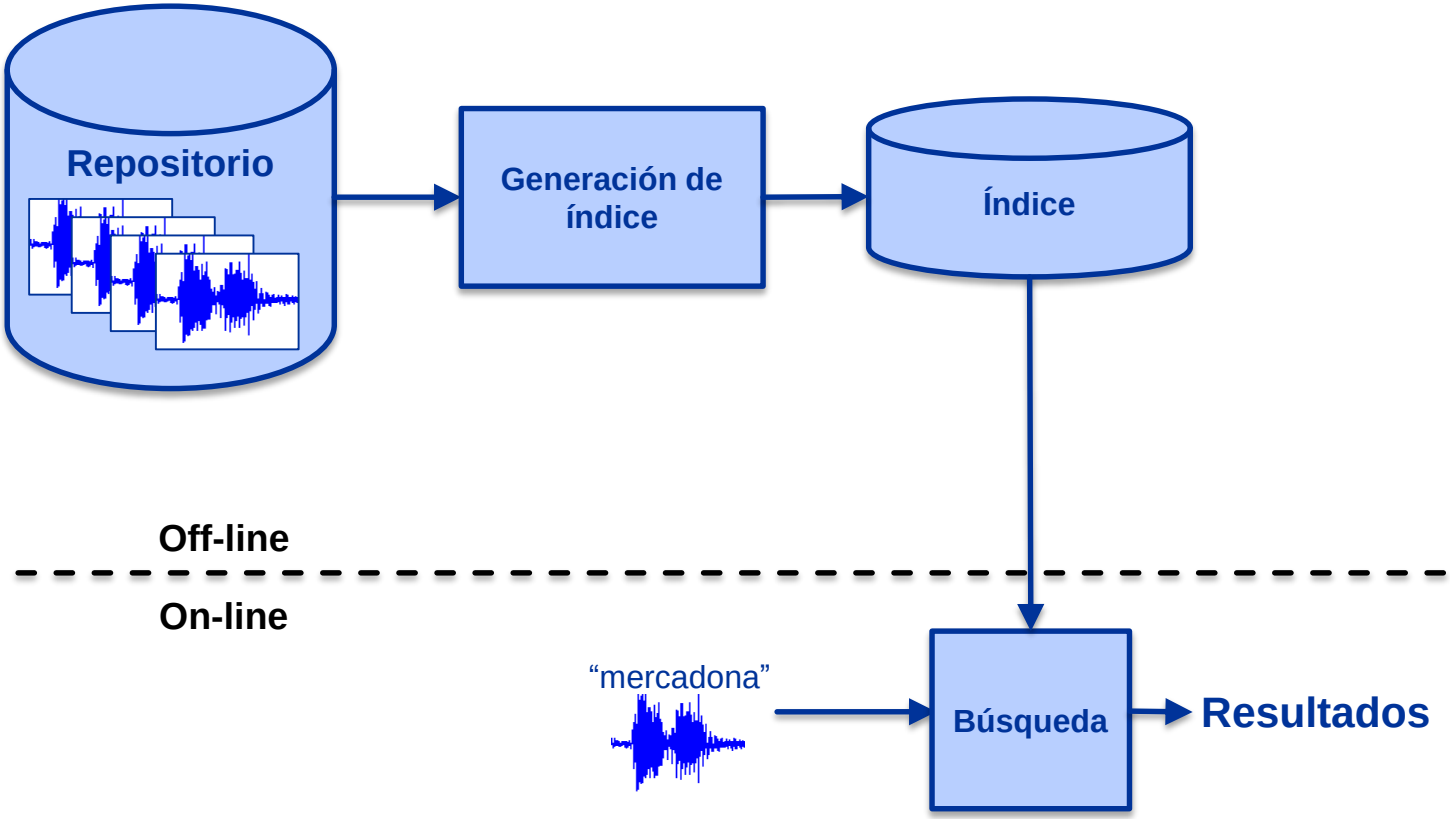
- Crecimiento de contenido audiovisual en repositorios multimedia y de audio
  - Archivos de TV y TV bajo demanda
    - Ejemplo: RTVE a la carta y archivo RTVE (más de 1000 horas)
  - Grabaciones de call centers
    - Por motivos legales, para control de calidad, para evaluación de campañas, para análisis de datos de clientes,…
  - Grabaciones de agencias de seguridad
    - Intercepción legal de comunicaciones
  - Grabaciones de interacción con dispositivos móviles
    - Smartphones, altavoces inteligentes, …
  - Grabaciones en el entorno judicial
    - Grabaciones de vistas orales

Motivación

- Los buscadores de texto permiten manejar eficientemente enormes colecciones de documentos
- Pero todavía no existen buscadores tan eficaces para buscar **en el contenido multimedia**
- Incluso si se ha encontrado un documento prometedor, para encontrar lo buscado
  - Muchas veces hay que ver o escuchar el contenido completo
  - El contenido completo puede ser de varias horas
  - Y no se puede ver/escuchar mucho más deprisa de tiempo real
- ¿No podríamos conseguir con **contenidos multimedia** la misma eficiencia y precisión que se consigue con documentos textuales?

Concepto de búsqueda en voz

- Localizar información en contenidos multimedia a partir de *queries* de texto o de voz

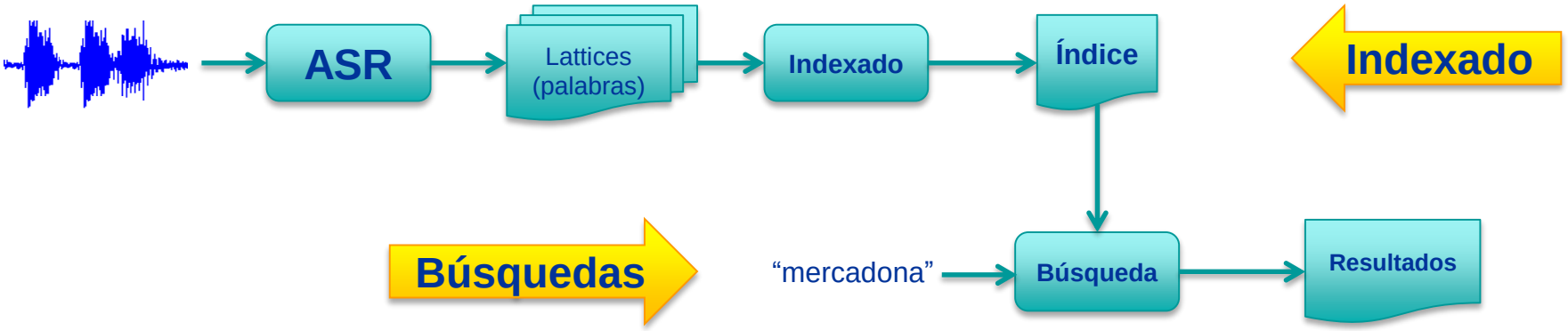


Tipos de búsquedas en voz

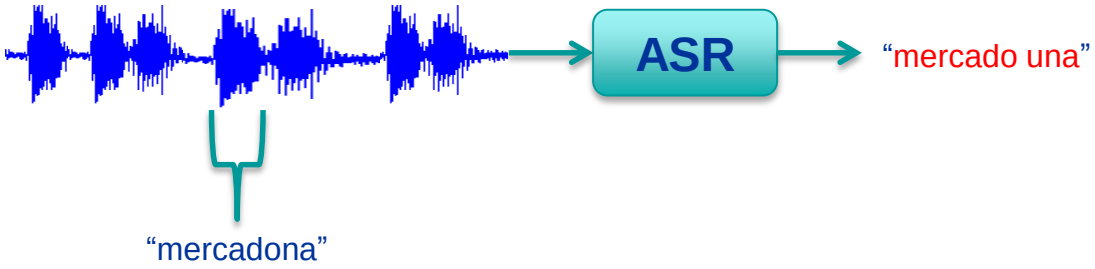
		Resultado	
		Lista documentos	Lista de coincidencias indicando documento + tiempo de cada una
Tipo de Query	Textual	Spoken Document Retrieval (SDR)	Spoken Term Detection (STD) o Keyword Spotting (KS)*
	Voz	Query-by-Example Spoken Document Retrieval (QbE SDR)	Query-by-Example Spoken Term Detection (QbE STD)

* STD y Keyword spotting se diferencian en que en STD se procesa el audio una vez y se genera un índice sobre el que posteriormente se puede buscar cualquier palabra.

Spoken Term Detection (STD): Un esquema típico

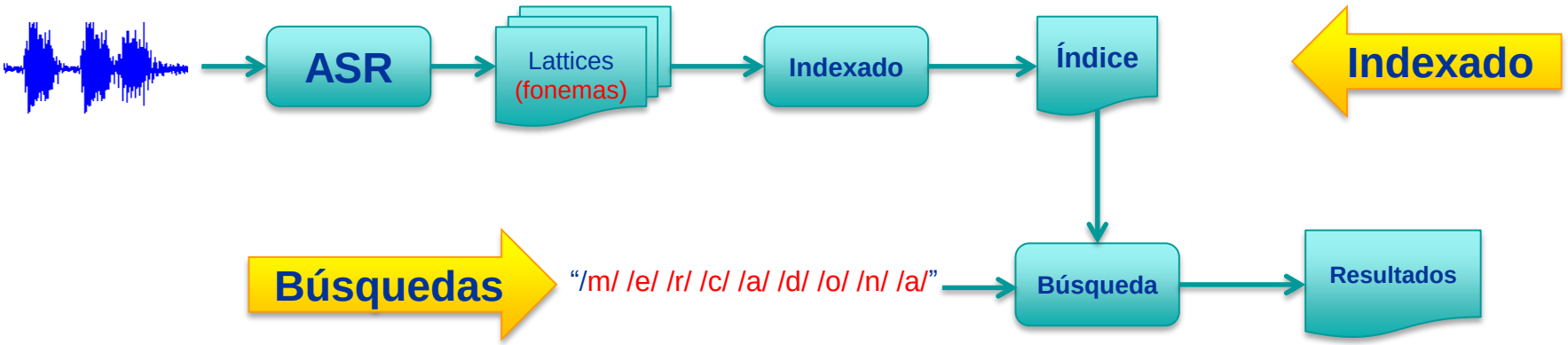


- Problema: Palabras fuera-de-vocabulario (OOVs) (10% de las palabras en una query)



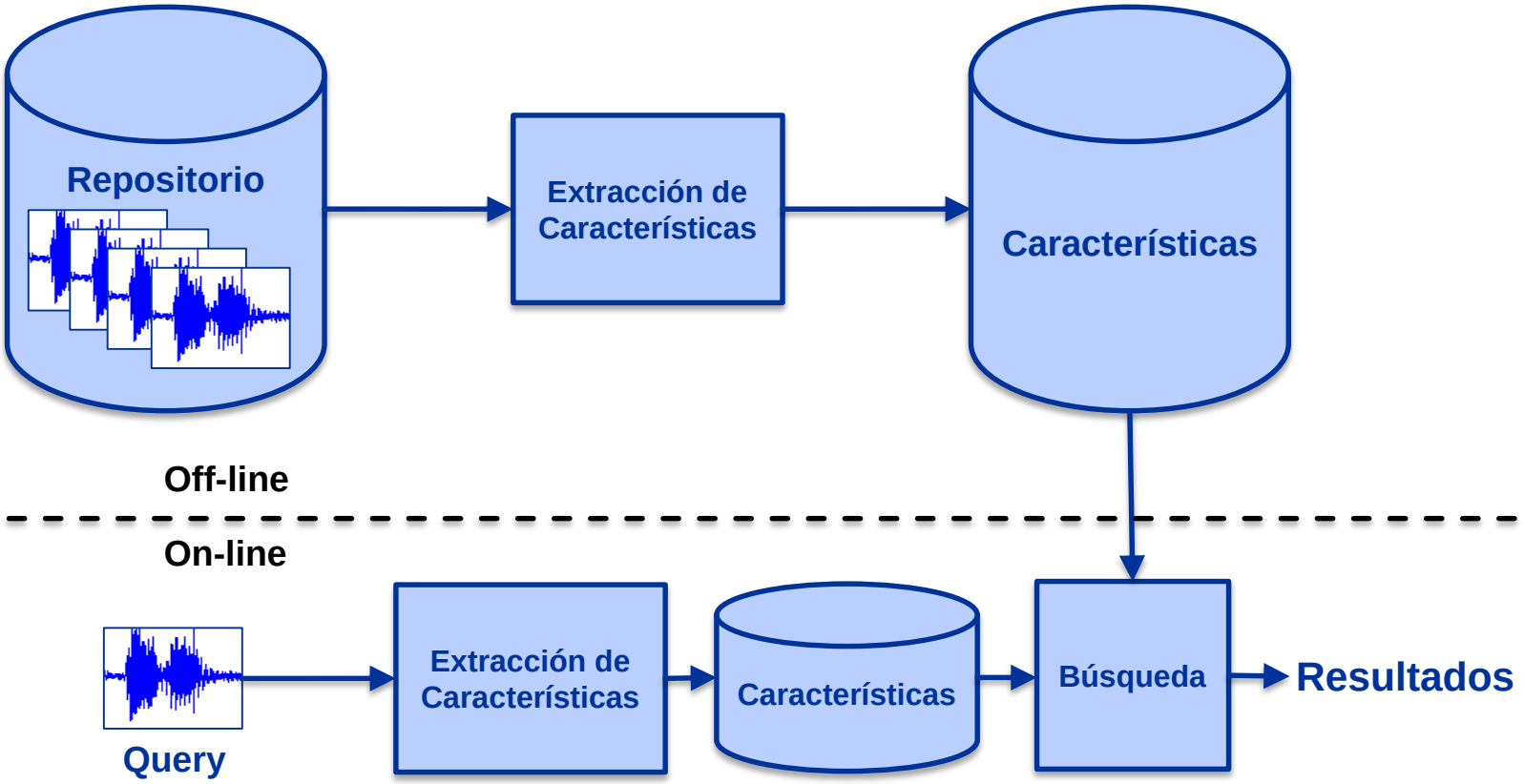
STD: Posibles soluciones

- Búsqueda en sub-unidades de palabra: fonemas, grafemas, sílabas,…



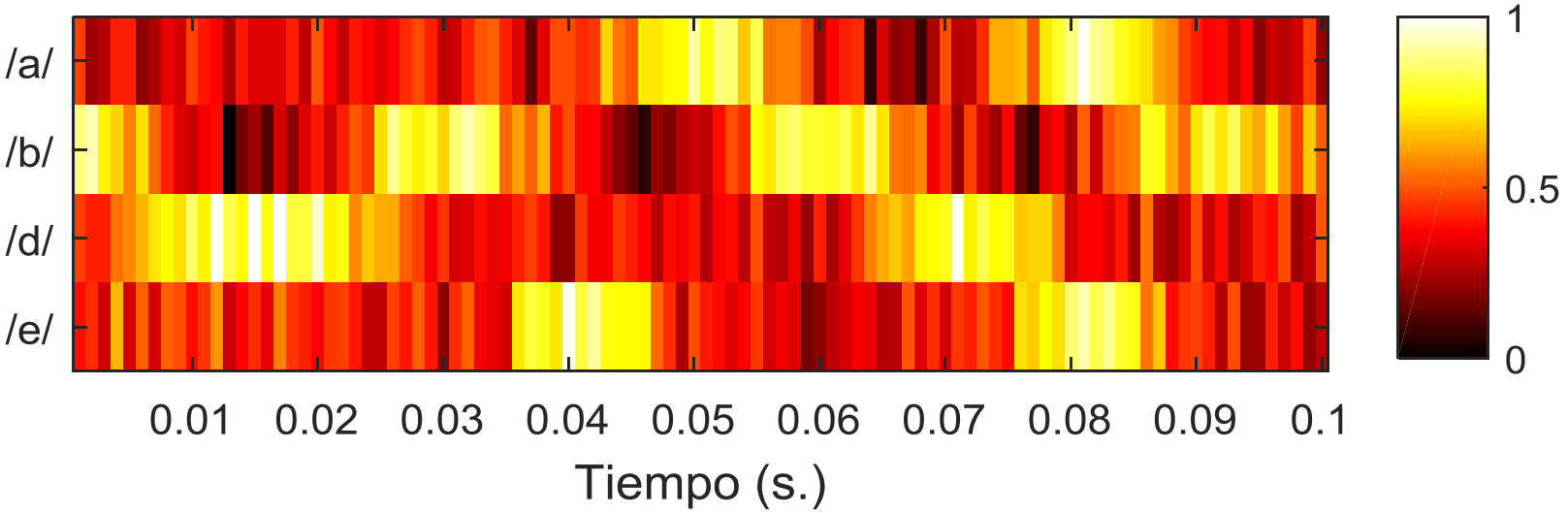
- Proxy-words: Usar el esquema típico y sustituir la palabra a buscar por la/s que más se parezca/n en el diccionario del ASR "mercado una" -> "mercadona"
- Solución final: Combinación de esquema típico+posibles soluciones.

Query-by-Example (QbE) STD: Un esquema típico



QbE STD: Extracción de Características

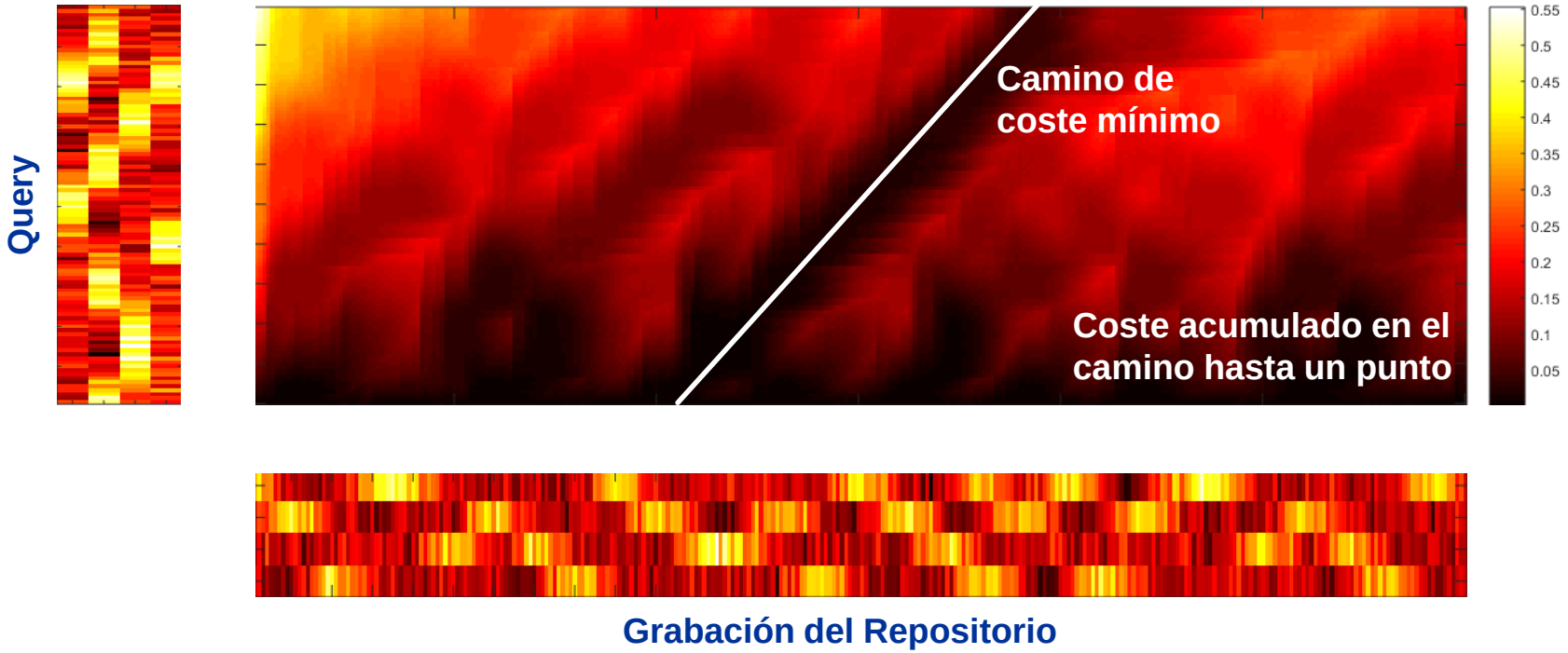
- Las más típicas son los *posteriorgramas* fonéticos:
  - Probabilidad de cada fonema en cada instante de tiempo
  - Obtenidas con un reconocedor de voz especial (sólo fonemas)



- Otras alternativas:
  - Características acústicas, Bottleneck features (BNF), etc.

QbE STD: Búsqueda

- Lo más habitual es la búsqueda en el repositorio de sub-secuencias de características similares a las extraídas de la *query*
  - *Subsequence Dynamic Time Warping (S-DTW)*



QbE STD: Casos de uso y ventajas

- Casos de uso:
  - Pronunciamos el contenido a buscar
    - Útil en dispositivos sin teclado (ej. smartphones, altavoces inteligentes, …)
  - Hemos encontrado un contenido interesante en un repositorio, lo seleccionamos y buscamos otros similares
- Ventajas:
  - Independencia del idioma y del tipo de lenguaje
  - No existe el problema de las OOVs
- Inconvenientes:
  - Menor precisión, en general
  - Búsquedas más lentas

Evaluación de sistemas de búsqueda en voz

- Con un conjunto de evaluación que incluya
  - Repositorio de documentos con voz
    - Mínimo 2 horas
  - Conjunto de *queries* (de texto o voz) a buscar
    - Que aparezcan suficientes veces
    - Algunas que no aparezcan
  - Etiquetado de las *queries* en los documentos
    - Todas en todos los documentos
    - Con su tiempo inicial y final

Evaluación de sistemas de búsqueda en voz

- Se clasifican las detecciones del sistema en:
  - Aciertos → *hits* (H)
  - Errores → falsas alarmas (FA)
  - En ambos casos, con cierta tolerancia en los tiempos (ej. 0.5s)
- ATWV: Actual Term Weighted Value
  - Combinación ponderada de H y FA promediada por *query*
    - ATWV mayor → mayor precisión
    - Máximo en 1, sin mínimo (puede ser negativa)
  - Depende del umbral de decisión
    - Buen sistema con mal umbral → ATWV bajo
- MTWV: Maximum Term Weighted Value
  - Máximo ATWV para todos los umbrales posibles
  - Mide bondad del sistema, sin tener en cuenta umbral

Evaluaciones competitivas de búsqueda en voz

- Permiten analizar el avance de cierta tecnología

Nombre	Años	Idioma
NIST TREC	1997-1999	Inglés
NIST STD	2006	Inglés, Árabe, Chino
NTCIR STD y QbE STD	2009-2016	Japonés
MediaEval	2011-2015	Inglés, Euskera, Checo, Rumano, Hindi, Telugu, etc
NIST OpenKWS	2013-2017	Vietnamita, Tamil, Swahili, Georgiano, Pashto e Inglés
NIST OpenSAT	2017-2019	Pashto, ??

- **Pero ninguna en castellano!**

Evaluaciones competitivas de búsqueda en voz

- Así que decidimos organizarlas nosotros:
  - ALBAYZIN Search-on-Speech 2012, 14, 16 y 18 en castellano:
    - Diferentes modalidades de entrada: texto y voz (STD y QbE STD)
    - Diferentes dominios de búsqueda (conferencias, noticias y conversaciones)
  - Basándonos en lo aprendido participando en:
    - MediaEval QbE STD 2011 (Inglés, Hindi, Gujarati, Telugu)
    - NIST Keyword Search STD 2013 (Vietnamita)
    - Muchas otras evaluaciones (NIST, ALBAYZIN) previas

Evaluaciones ALBAYZIN SoS 2012 – 2018

Año	Participantes (Sistemas)	Dominio	Nº horas	Nº queries Qbe STD (oc)	Nº queries STD (oc)
2012	5 (10)	Conferencias	2	60 (892)	155 (1971)
2014	7 (10)	Conferencias	2	193 (2577)	548 (6246)
2016	10 (18)	Conferencias	4.5	303 (2769)	780 (5047)
2018	6 (24)	Conferencias Noticias Conversaciones	28.5	510 (3928)	1560 (6919)

Año	MTWV (STD)	ATWV (STD)	MTWV (QbE STD)	ATWV (QbE STD)
2012	0.0246	0.0043	0.0217	0.0217
2014	0.5451	0.5350	0.2708	0.2708
2016	0.5850	0.5724	0.2739	0.2646
2018	TBA	TBA	TBA	TBA

Conclusiones

- Búsqueda en voz necesaria en la era digital
- Un sistema de ASR no soluciona por completo (aunque ayuda y mucho) la problemática de la búsqueda en voz
- Los sistemas están lejos de ser perfectos
- El dominio de los datos juega un papel fundamental a la hora de analizar el rendimiento del sistema
- **Las evaluaciones son esenciales para avanzar y medir el avance**

Agradecimientos

- Red Temática en Tecnologías del Habla (RTTH)
- Laboratorio de Lingüística Informática Universidad Autónoma de Madrid
- Cátedra RTVE-Universidad de Zaragoza
- Real Academia de Ingeniería (RAI)
- Universidad Autónoma de Madrid
- Universidad CEU San Pablo