



# Indexación de audio:

## Tema 1: Introducción

< audias >

Audio, Data Intelligence and Speech

<http://audias.ii.uam.es>

Daniel Ramos Castro

Contrib. de Doroteo Torre, Joaquín González, Alicia Lozano

1

< audias >

Audio, Data Intelligence and Speech

Home News Seminars People Publications Projects & Industry Evaluations For Students Contact



Home



AUDIAS is a solid research group addressing challenging problems in **speech, audio and temporal signals** from deep foundations in **machine learning and signal processing**.

With a strong focus in **technology transfer**, AUDIAS has contributed for more than 20 years to the progress of speech and audio technologies and forensic science.

AUDIAS, which continues the speech and audio activities of former ATVS research group, **is currently focused in:**

- Speech and language technologies.
- Audio and music analytics.
- Data science in signals from industrial, vehicle and biomedical sensors.
- Data intelligence in financial series.

### AUDIAS Seminars

[Perceiver: General Perception with Iterative Attention](#)

© October 7, 2022

Speaker: Juan Ignacio Álvarez Trejos.

Abstract: Biological systems perceive the...

[Continual learning for recurrent neural networks](#)

© September 30, 2022



## People

< audias >

- AUDIAS = Audio, Data Intelligence and Speech
- Personal
  - ▣ 4 Profesores (2 Catedráticos, 1 Prof. Titular, 1 Prof. Ayte. Doctora)
  - ▣ 6 doctorandos (Ingenieros + Master)
  - ▣ ~ 5-10 estudiantes de Grado y Máster

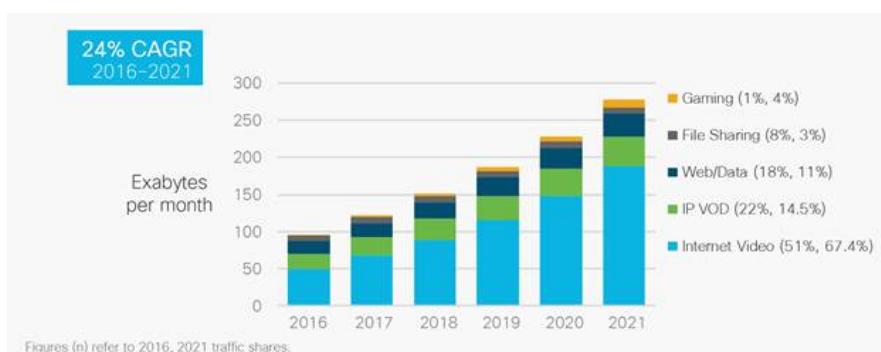


Máster en Big Data y Data Science

Indexación de voz & audio

## 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



Máster en Big Data y Data Science

Indexación de voz & audio

3

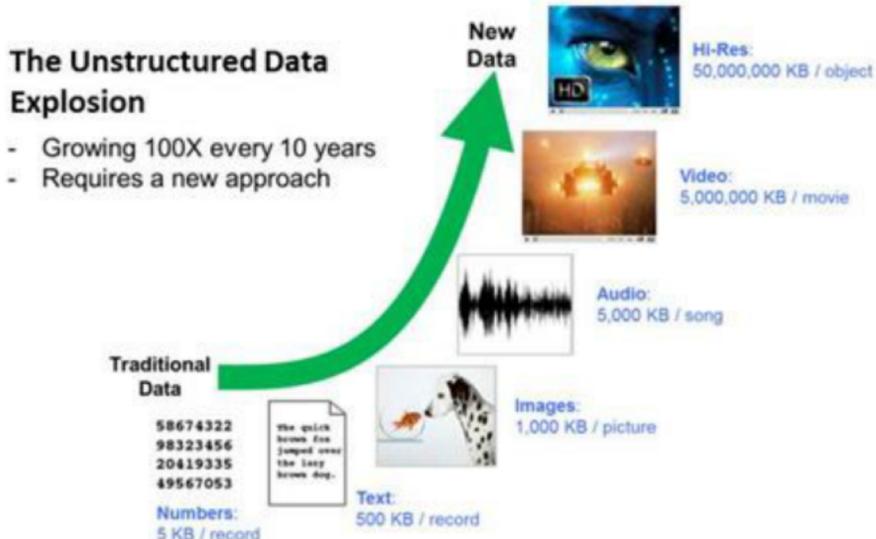
## Motivación

- 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?

## Tipos de datos No Estructurados



Audio “in the wild”: welcome to the jungle!



## Audio “in the wild”: welcome to the jungle!

- Microphone, mobile, media, ATC (pilot-control)



- Internet videos



- Emergency recordings (firefighters in action)

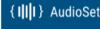


## Google AudioSet

### A sound vocabulary and dataset

AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds.

By releasing AudioSet, we hope to provide a common, realistic-scale evaluation task for audio event detection, as well as a starting point for a comprehensive vocabulary of sound events.

 AudioSet

HOME   ONTOLOGY   DATASET   DOWNLOAD   ABOUT

**Ontology**

The AudioSet ontology is a collection of sound events organized in a hierarchy. The ontology covers a wide range of everyday sounds, from human and animal sounds, to natural and environmental sounds, to musical and miscellaneous sounds.

Each ontology entry contains a description of the sound event, modified from [Wikipedia](#), [WordNet](#), or written by us. The ontology is meant to be expandable to meet to research needs of the academic community. Visit our [GitHub](#) repository to download and contribute to our growing collection of audio event knowledge.

You can learn more about the construction of the ontology in our [ICASSP 2017 paper](#). Explore the sound classes in the ontology below.

Type a sound to filter the ontology

<b>Human sounds</b> <ul style="list-style-type: none"> <li>— Human voice</li> <li>— Whistling</li> <li>— Respiratory sounds</li> <li>— Human locomotion</li> <li>— Digestive</li> <li>— Hands</li> <li>— Heart sounds, heartbeat</li> <li>— Otacoustic emission</li> <li>— Human group actions</li> </ul> <b>Source-ambiguous sounds</b> <ul style="list-style-type: none"> <li>— Generic impact sounds</li> <li>— Surface contact</li> <li>— Deformable shell</li> <li>— Onomatopoeia</li> <li>— Silence</li> <li>— Other sourceless</li> </ul>	<b>Animal</b> <ul style="list-style-type: none"> <li>— Domestic animals, pets</li> <li>— Livestock, farm animals, working animals</li> <li>— Wild animals</li> </ul> <b>Sounds of things</b> <ul style="list-style-type: none"> <li>— Vehicle</li> <li>— Engine</li> <li>— Domestic sounds, home sounds</li> <li>— Bell</li> <li>— Alarm</li> <li>— Mechanisms</li> <li>— Tools</li> <li>— Explosion</li> <li>— Wood</li> <li>— Glass</li> <li>— Liquid</li> <li>— Miscellaneous sources</li> </ul>	<b>Music</b> <ul style="list-style-type: none"> <li>— Musical instrument</li> <li>— Music genre</li> <li>— Musical concepts</li> <li>— Music role</li> <li>— Music mood</li> </ul> <b>Natural sounds</b> <ul style="list-style-type: none"> <li>— Wind</li> <li>— Thunderstorm</li> <li>— Water</li> <li>— Fire</li> </ul> <b>Channel, environment and background</b> <ul style="list-style-type: none"> <li>— Acoustic environment</li> <li>— Noise</li> <li>— Sound reproduction</li> </ul>
---	--	---

**Máster en Big Data** **exación de voz & audio**

**EXPLORE THE ONTOLOGY**

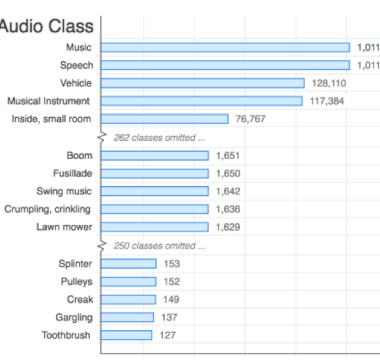
<b>2.1 million</b> annotated videos	<b>5.8 thousand</b> hours of audio	<b>527 classes</b> of annotated sounds
--	---------------------------------------	---

**Large-scale data collection**

To collect all our data we worked with human annotators who verified the presence of sounds they heard within YouTube segments. To nominate segments for annotation, we relied on YouTube metadata and content-based search.

Our resulting dataset has excellent coverage over the audio event classes in our ontology.

**Audio Class**



Audio Class	Number of examples
Music	1,011,949
Speech	1,011,065
Vehicle	128,110
Musical Instrument	117,384
Inside, small room	76,767
Boom	1,651
Fusillade	1,650
Swing music	1,642
Crumpling, crinkling	1,636
Lawn mower	1,629
Splinter	153
Pulleys	152
Creak	149
Gargling	137
Toothbrush	127

Number of examples

**Máster en Big Data** **udio**

# Ejemplos de Algunas Aplicaciones

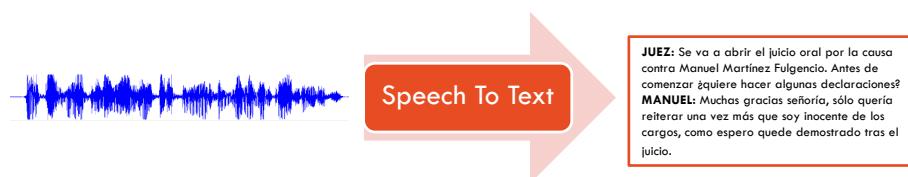
< audias >

Audio, Data Intelligence and Speech  
<http://audias.ii.uam.es>

Daniel Ramos Castro

Contrib. de Doroteo Torre, Joaquín González, Alicia Lozano

## Conversión Voz a Texto (STT)



< audias >

Audio, Data Intelligence and Speech

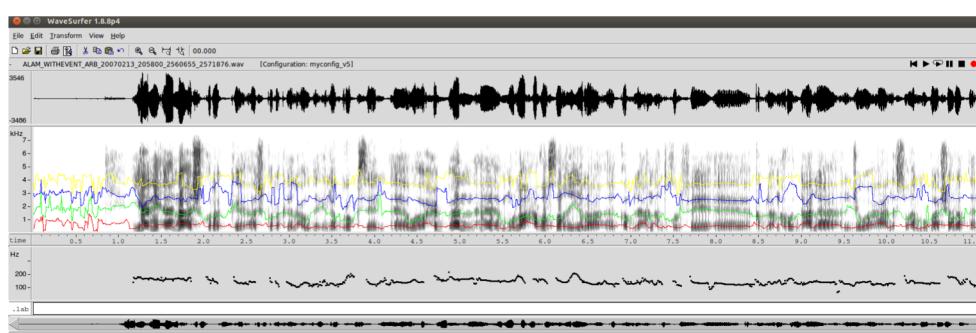
UAM

## Reconocimiento de voz en Árabe

- Base de Datos: GALE Arabic – habla conversacional en medios de comunicación recogida entre 2006 y 2007 por LDC como parte del programa DARPA GALE (Global Autonomous Language Exploitation)
- Voz Conversacional y de Reportajes
- Datos de entrenamiento: 320 horas
- Datos de test: 9.3 horas
- Sistema desarrollado: TDNN, 7 capas ocultas de 850 unidades (ReLU), uso de trigrafemas (2968)
- Métricas:
  - WER (Word Error Rate): 12.24% (reportaje), 27.73% (conversacional), 22.76% (combinado)



## Ejemplos de Reconocimiento de voz en Árabe





## Air Traffic Control (ATC) STT (Inglés)

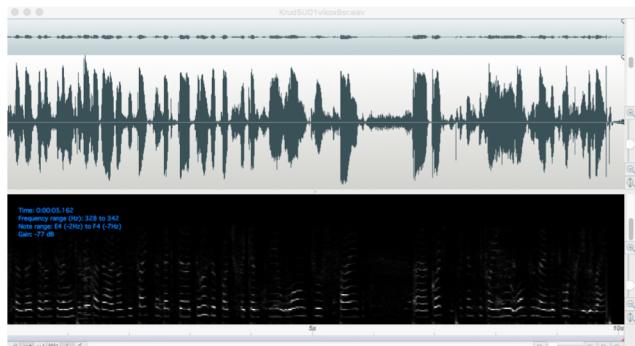
- Desarrollado para la evaluación AIRBUS ATC ASR Challenge 2018
- Base de Datos: subconjuntos del corpus de voz AIRBUS-ATC
  - Entrenamiento: 40 horas, Desarrollo (leaderboard): 5 horas, Test: 5 horas
- Sistema Desarrollado:
  - TDNN, 6 capas ocultas de 1024 unidades, capa de salida de 4440 unidades.
  - Entrenamiento con datos aumentados (perturbaciones de velocidad)
  - I-vector con información del hablante añadido al vector de características para compensar las características de cada locutor
  - Léxico generado a partir de la base de datos de entrenamiento y ampliado con palabras de nombres de aerolíneas y torres de control
  - Conversión grafema-fonema (G2P) basada en LSTMs secuencia-a-secuencia entrenados con el léxico de Switchboard
  - Modelo de lenguaje (trígrama) entrenado sólo con las transcripciones del corpus de entrenamiento
- Resultados: WER 11.36%, F1 score (0-1): 0.72



### Call signs:

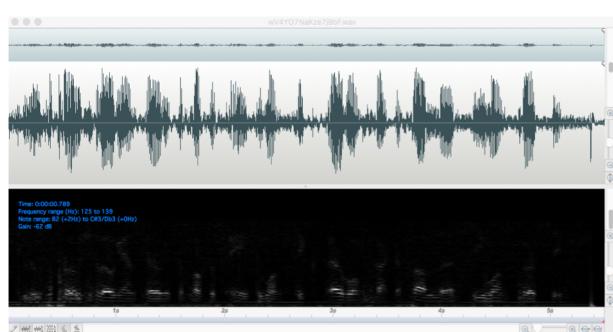
- Real: "Lufthansa zero two juliet"
- Detectado: "lufthansa zero two juliett november"

- Transcripción: "Lufthansa zero two juliett november four cross three two right maintain holding point mike four three two left"
- Reconocido: "lufthansa zero two juliett november four cross three two right maintain holding point mike four three two left"



Call signs:

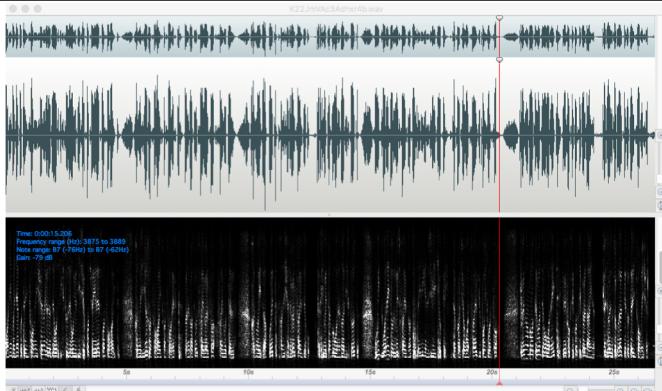
- Real: "Lufthansa seven seven xray"
- Detected: "lufthansa seven seven xray"
- Transcripción: "Lufthansa seven seven xray radar identified contact Toulouse approach frequency one two five huh correction frequency for you one two nine decimal three goodbye"
- Reconocido: "lufthansa seven seven xray radar identified contact toulouse approach frequency one two five huh correction frequency for you one two nine decimal three goodbye"



Call signs:

- Real: "sierra victor foxtrot eight one six"
- Detected: "sierra victor foxtrot eight one six"

- Transcripción: "Toulouse approach hello again sierra victor foxtrot eight one six arrival \_ six one thirty"
- Reconocido: "toulouse approach hello again sierra victor foxtrot eight one six arrival six one thirty"



#### Call signs:

- Real: "(vacío)"
- Detectado: "(vacío)"

- Transcripción: / morning Blagnac hotel information recorded at zero six two zero UTC approach ILS one four right / runways in use one four right one four left \_ planned departure route five alpha five hotel / runways are wet transition level five zero / taxiway tango one hundred is closed / wind one eight zero degrees five knots ceiling and visibility okay / temperature one five dew point one four / quebec november hotel one zero one eight quebec fox echo triple nine / advised at first contact you got hotel information
- Reconocido: morning blagnac hotel information recorded at zero six two zero utc approach ils one four right runways in use one four right one four left planned departure route five alfa five hotel runways are wet transition level five zero taxiway tango one hundred is closed wind one eight zero degrees five knots ceiling and visibility okay temperature one five dewpoint one four quebec november hotel one zero one eight quebec fox echo triple nine advised at first contact you got hotel information

## Clasificación de Tipos de Audio

< audias >  
Audio, Data Intelligence and Speech

UAM

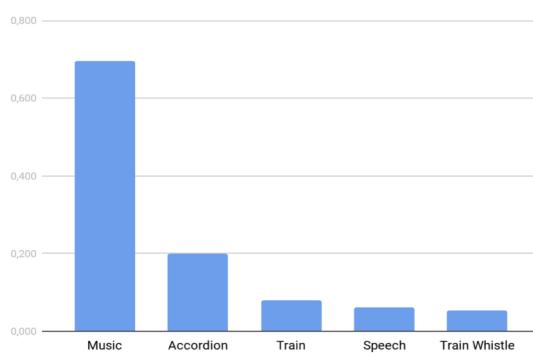
## Google AudioSet (2017)

- ~1.8M YouTube videos
- Audio clips of 10 seconds = total of ~5000 hours
- Audio content annotated by humans
  - Presence, not timing
- 7 classes & 527 subclasses
- Experimental setup:
  - Training set (unbalanced): 1.771.873 clips
  - Test set: 17.748 clips
- Addressed by Audias through DNNs, CNNs & LSTMs

- Human sounds
  - Human voice
  - Whistling
  - Respiratory sounds
  - Human locomotion
  - Digestive
  - Hands
  - Heart sounds, heartbeat
  - Oraacoustic emission
  - Human group actions
- Sounds of things
  - Vehicle
  - Engine
  - Domestic sounds, home sounds
  - Bell
  - Alarm
  - Mechanisms
  - Tools
  - Explosion
  - Wood
  - Glass
  - Liquid
  - Miscellaneous sources
  - Specific impact sounds
- Animal sounds
  - Domestic animals, pets
  - Livestock, farm animals, working animals
  - Wild animals
- Natural sounds
  - Wind
  - Thunderstorm
  - Water
  - Fire
- Music
  - Musical instrument
  - Music genre
  - Musical concepts
  - Music role
  - Music mood
- Source-ambiguous sounds
  - Generic impact sounds
  - Surface contact
  - Deformable shell
  - Onomatopoeia
  - Silence
  - Other sourceless
- Channel, environment and background
  - Acoustic environment
  - Noise
  - Sound reproduction

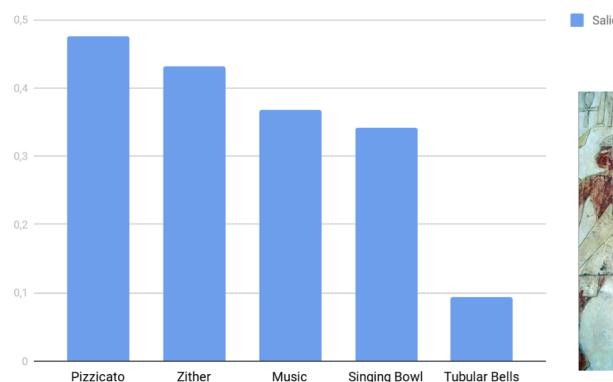
## Sample videos & LSTM classification

- Correct (labeled) class: “Accordion”



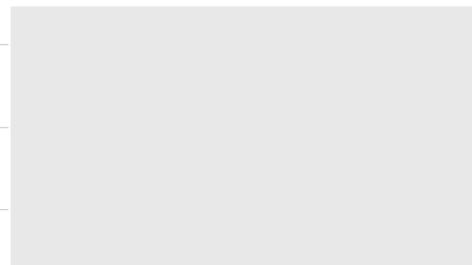
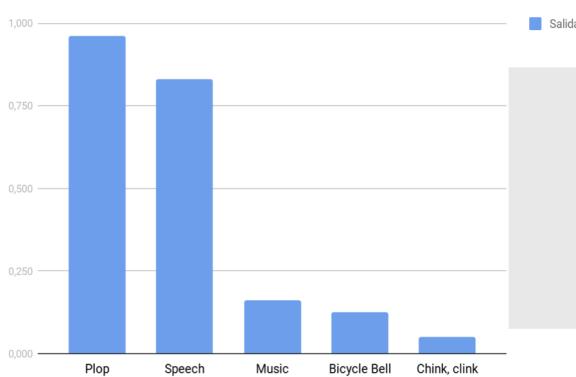
## Sample videos & LSTM classification

- Correct class: “Music” & “Traditional Music”



## Sample videos & LSTM classification

- Correct class: “Speech” & “Plop”



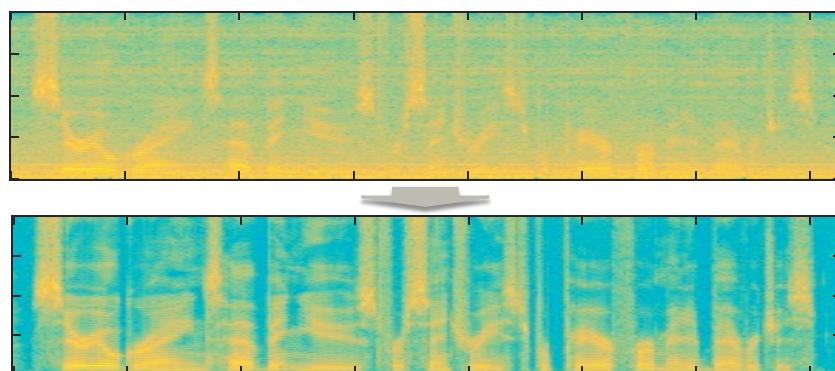
## Mejora de Voz con DNNs



Máster en Big Data y Data Science

Indexación de voz & audio

## Objetivo



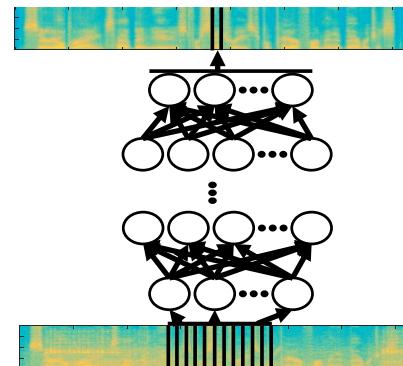
Máster en Big Data y Data Science

Indexación de voz & audio

29

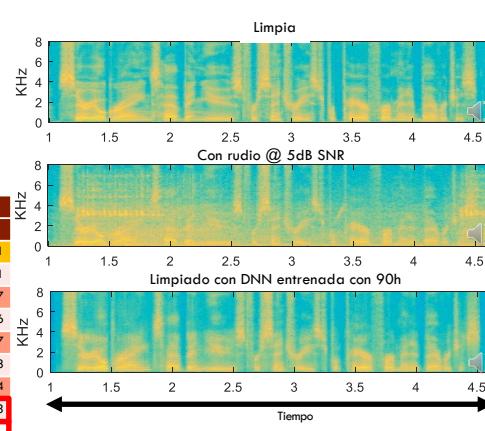
## Aproximación: Entrenar una DNN que aprenda a limpiar espectros de voz

- Entrada:
  - Frame ruidoso (Central)
  - Entorno ruidoso (Central  $\pm N$  frames ruidosos)
  - Voz ruidosa sintética
- Salida:
  - Frame limpiado (Central)
- Target (entrenamiento)
  - Frame limpio (central)



## Ejemplo y Evaluación

- Test data:
  - TIMIT Core Test Set
  - HU Noises



Ave. MOS	SNR				
	-5	0	5	10	15
Noisy speech	1,3633	1,5754	1,8043	1,6646	2,3151
Cleaned 1h	1,3953	1,6203	1,8219	1,6231	2,1211
2h	1,5484	1,7774	1,9755	1,7344	2,2777
5h	1,7773	2,0054	2,1916	1,8857	2,4786
10h	1,7293	1,9961	2,2126	1,8729	2,5277
20h	1,8583	2,1187	2,3243	1,951	2,6328
50h	1,9894	2,2363	2,4356	1,9979	2,7414
90h	2,0191	2,277	2,4808	2,0212	2,7853

34,3% mejora

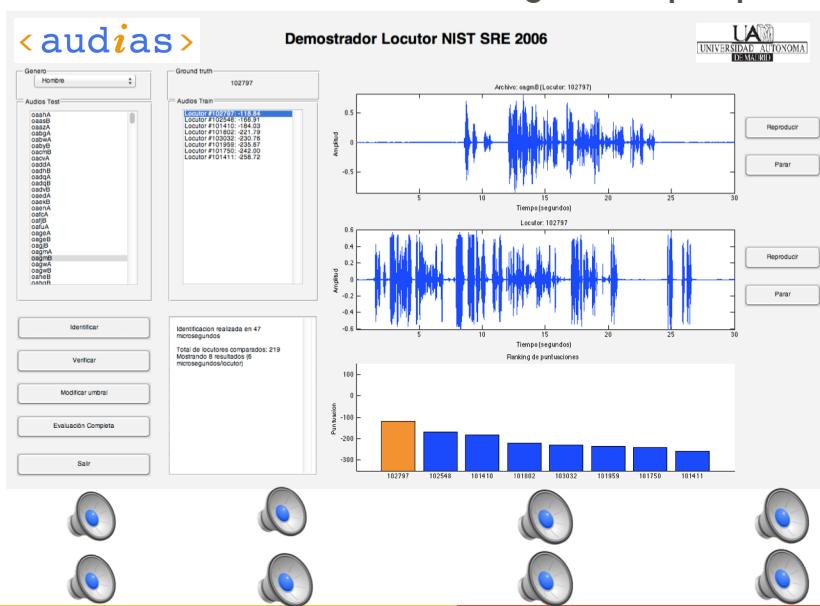
## Reconocimiento de Locutores



Máster en Big Data y Data Science

Indexación de voz &amp; audio

### Reconocimiento de Locutores: Algunos Ejemplos



Máster en Big Data y Data Science

Indexación de voz &amp; audio

33

## Reconocimiento de Idioma



Máster en Big Data y Data Science

Indexación de voz & audio

### Reconocimiento de Idioma: Algunos Ejemplos

**Demostrador Idioma NIST LRE 2009 (30 segundos)**

**Verificación de Idioma**

Lenguaje	Speaker 1	Speaker 2
Inglés	Speaker 1	Speaker 2
Castellano	Speaker 1	Speaker 2
Dari	Speaker 1	Speaker 2
Frances	Speaker 1	Speaker 2
Pashto	Speaker 1	Speaker 2
Ruso	Speaker 1	Speaker 2
Urdu	Speaker 1	Speaker 2
Mandarin	Speaker 1	Speaker 2

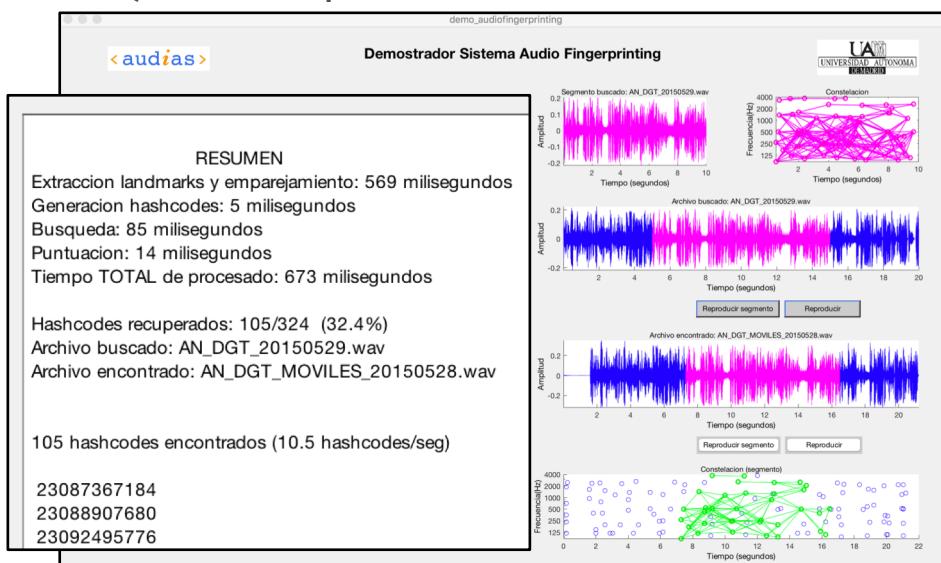
Archivo de audio: zzz.wav

Ranking de puntuaciones

Lenguaje	Puntuación
Pashto	~0.15
Dari	~0.07
Frances	~0.06
Russo	~0.03
Mandarin	~0.02
Castellano	~-0.08
Ingles	~-0.12
Urdu	~-0.15

5

## Audio Fingerprinting: Búsqueda de un Fragmento (Audio Enlatado) en un Repositorio



Máster en Big Data y Data Science

Indexación de voz &amp; audio

36



Máster en Big Data y Data Science

Indexación, búsqueda y análisis  
en repositorios multimedia

## Señal de Audio Digital

< audias >

Audio, Data Intelligence and Speech

<http://audias.ii.uam.es>

Daniel Ramos Castro

Contrib. de Doroteo Torre, Joaquín González, Alicia Lozano

## ¿Qué es en una señal de audio?

### ➤ Audio analógico

- Señal (típicamente de tensión eléctrica)
- Representa las variaciones de presión acústica debidas al sonido
- Se capta con micrófonos
- Se reproduce con altavoces
- Se representa en el tiempo como la llamada “forma de onda”:
  - Eje x tiempo
  - Eje y “amplitud” (puede ser tensión eléctrica en voltios, o valor sin dimensión)
  - Típicamente expresado en decibelios (dB)

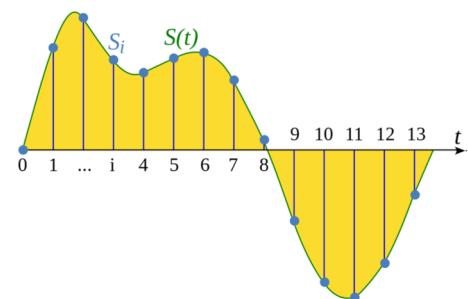


## Muestreo y cuantificación: Conversión Analógico a Digital

- Es importante poder manejar audio con un ordenador
- Para ello hay que transformar el **audio analógico** en **audio digital**
- ¿Por qué?
  - Un soporte digital (memoria, ordenador) no puede almacenar información infinita, sino finita
  - Objetivo: convertir la señal analógica (continuo, infinitos valores) en un conjunto de valores discretos (finitos)
  - Para poder guardar todo ello en un ordenador
- Procedimiento: conversión analógico-digital (CA/D, DAC en inglés), formado por dos procesos
  - Primer proceso: muestreo
  - Segundo proceso: cuantificación
  - Objetivo: convertir señal analógica en “unos y ceros” para guardar en un soporte digital

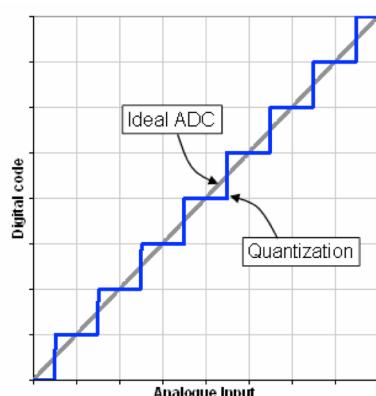
## CA/D: paso 1, muestreo

- Convierte la señal analógica ( $S(t)$ ) “infinitas” muestras en el tiempo, continuo)
- En muestras de la señal en tiempo discreto ( $S_i$ )
  - Parámetro esencial en este proceso: frecuencia de muestreo (Hz)
    - Define cuántas muestras se toman por segundo de la señal analógica
    - Inverso del periodo de muestreo (tiempo entre muestras)
- Valores típicos de frecuencia de muestreo:
  - 44,1 KHz: alta calidad, “formato compact-disc (CD)”
  - 8KHz: voz telefónica (calidad media, se persigue la inteligibilidad)



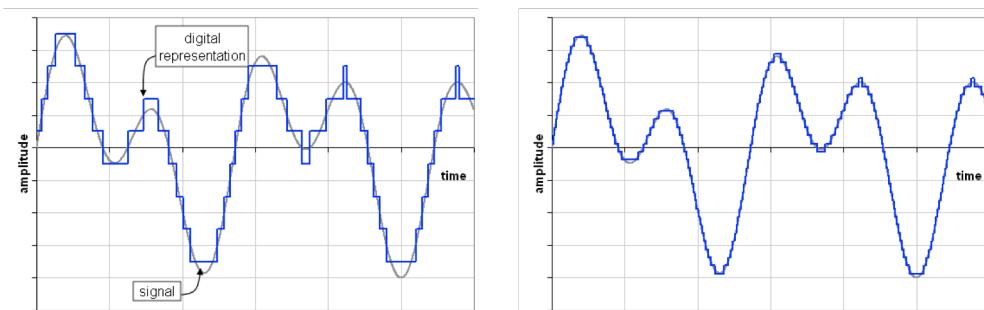
## CA/D: paso 2, cuantificación

- Las muestras en el tiempo siguen siendo continuas en su valor de amplitud
- Necesario convertir el rango de amplitud de señal en una serie de niveles discretos
- Eso lo hace un **cuantificador**
  - Número de niveles típicos: 216
    - Por tanto, se almacenan en 16 bits por cada muestra
      - “Profundidad” de cuantificación



## Resultado de CA/D

- La señal analógica (continua en tiempo, continua en amplitud)
- Se transforma en 16 bits (nivel de cuantificación) por cada muestra (tomada un número de veces por segundo definido por la frecuencia de muestreo)
  - Señal digital
  - Discreta en tiempo, discreta en niveles



## Fichero de audio digital

Header
001101001010001001 001001001010100101 011011010101100101 010010101 ... .... 000100110101010

### ➤ Cabecera del fichero:

- Nombre y localización
- Número de canales (1 mono / 2 estéreo)
- Número de bits / muestra
- Frecuencia de muestreo (8-48 kHz)
- Tamaño (~ duración)

### ➤ Contenido

- **Ceros y unos representando las muestras**
- Datos no estructurados
- Tiempo análisis = tiempo escucha (1:1)

## La “estructura” del sonido

- No hay estructura a bajo nivel (aparte de la forma en la que se empaquetan los bits, bytes, etc.)
  - Endless flow of audio samples ( $p(nT_s)$ ),  $T_s=1/f_s$  (sampling frequency)
- Estructura a bajo nivel: provocada por cómo se generan los sonidos
  - Disparos, portazos, sonidos de máquinas...
  - Producción sonora de instrumentos musicales
  - Voz humana (fonética)
- Estructura de alto nivel: caracteriza la secuencia de sonidos y los sonidos simultáneos
  - Ejemplos:
    - Música: tono, temperamento, melodía, armonía, ritmo, instrumentación, estilos...
    - Lenguaje: fonotáctica, sintaxis, gramáticas, vocabularios, dialecto, sociolecto...



Máster en Big Data y Data Science

Indexación de voz &amp; audio



Máster en Big Data y Data Science

Indexación, búsqueda y análisis  
en repositorios multimedia

## Representación de Audio y Extracción de Parámetros

< audias >

Audio, Data Intelligence and Speech

<http://audias.ii.uam.es>

Daniel Ramos Castro

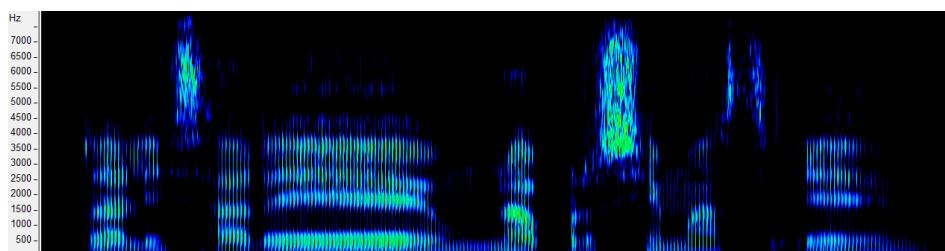
Contrib. de Doroteo Torre, Joaquín González, Alicia Lozano

## Introducción al procesamiento de audio y voz

- Normalmente no se procesa el audio o voz directamente

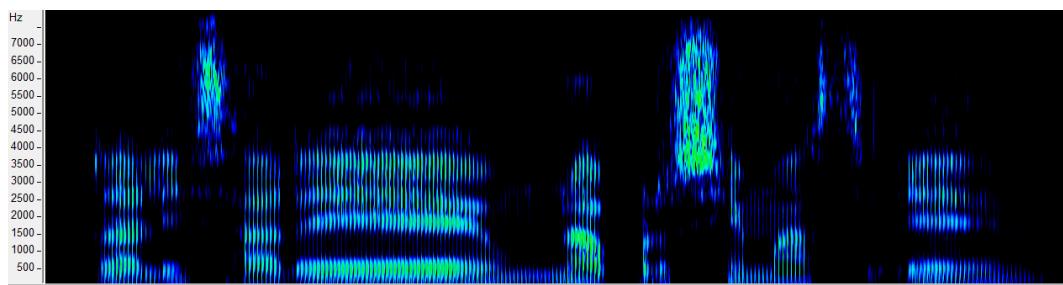


- Sino una representación del mismo:



## Introducción al procesamiento de audio y voz

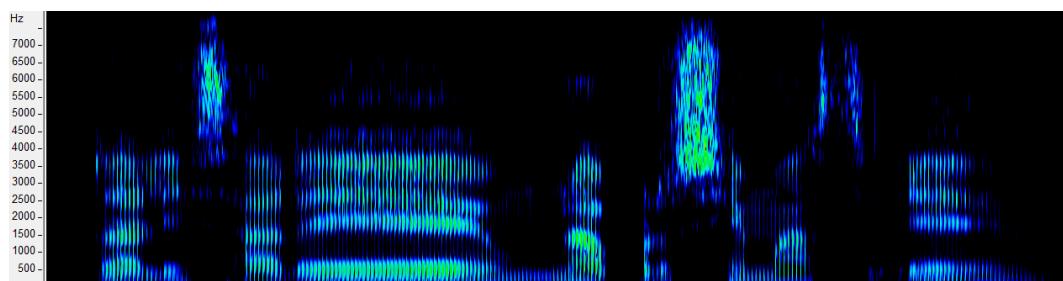
- Representación tipo “espectrograma”
- Eje x: tiempo (discreto)
- Eje y: espectro de la señal (transformada discreta de Fourier, módulo)



## Introducción al procesamiento de audio y voz

### ➤ Representación tipo “espectrograma”

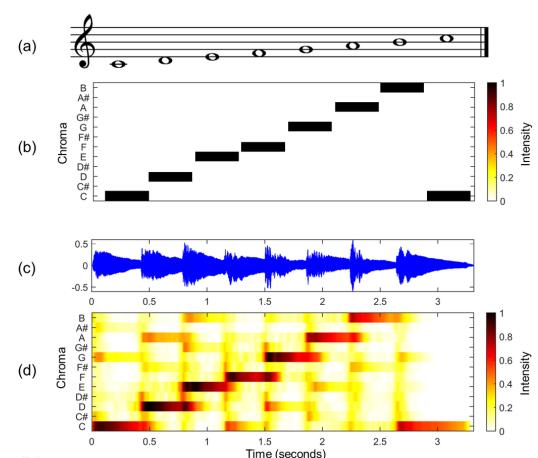
- Idea: cada tipo de sonido tiene un espectro diferente
  - Una vocal u otra vocal, música, sonido de aspiradora, etc.
  - En el spectrograma podemos distinguir esos sonidos
  - Eso permite una muy extensa variedad de tareas de indexación, búsqueda y análisis



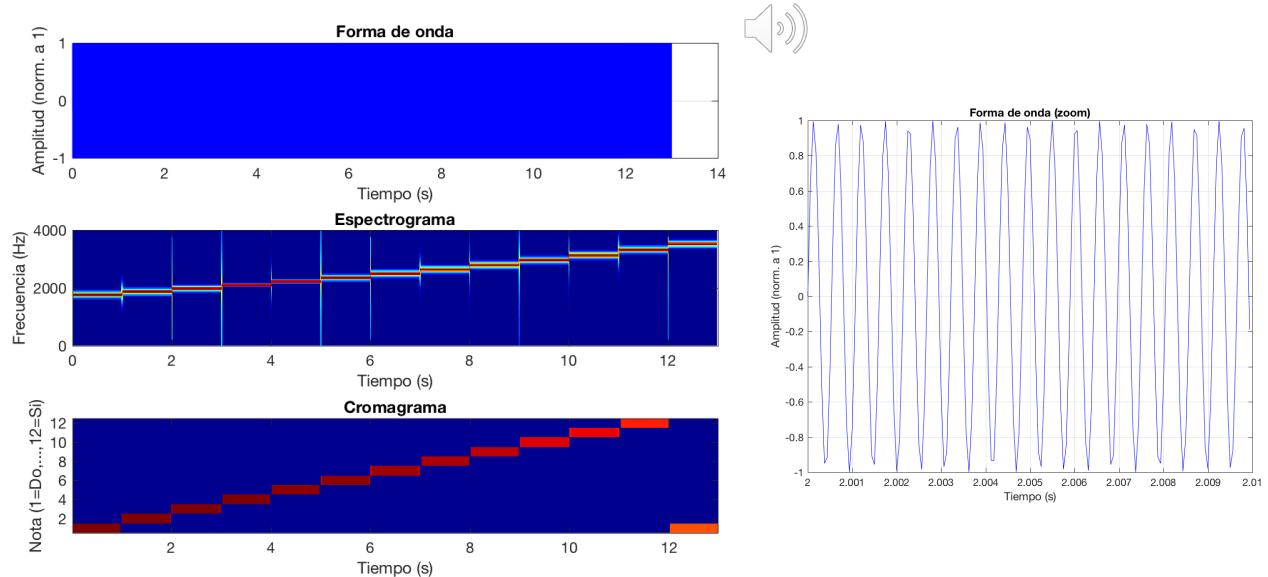
## Introducción al procesamiento de audio y voz

### ➤ Representación tipo “cromagrama”

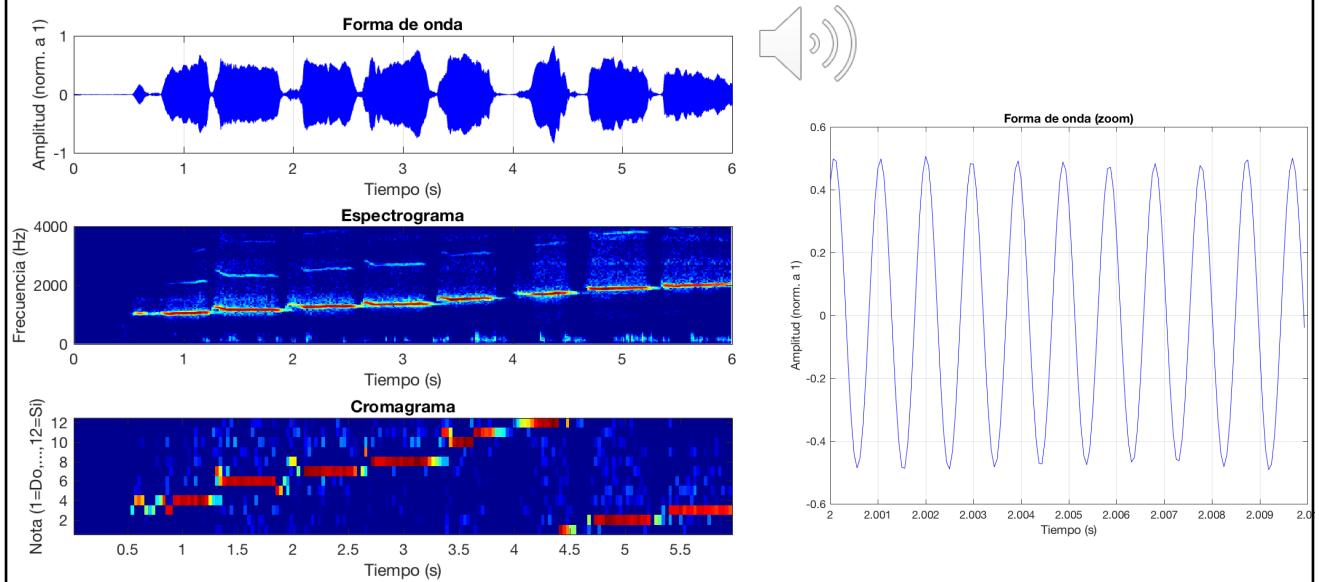
- Idea: cada tipo de sonido emite una serie de notas musicales
  - Una vocal u otra vocal, música, sonido de aspiradora, etc.
  - En el spectrograma podemos distinguir esos sonidos
  - Eso permite una muy extensa variedad de tareas de indexación, búsqueda y análisis



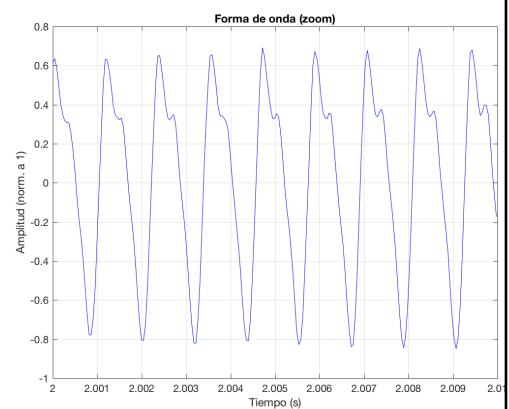
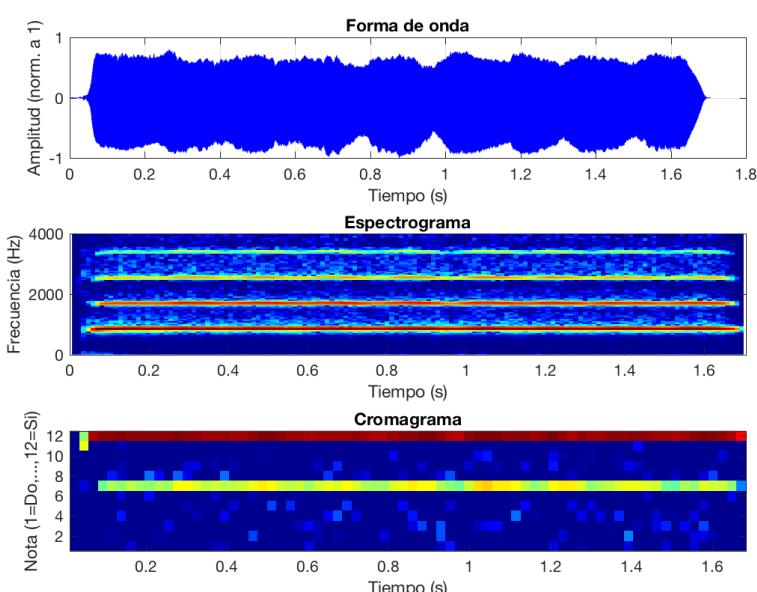
## Ejemplos de Representación de Señales de Audio



## Ejemplos de Representación de Señales de Audio



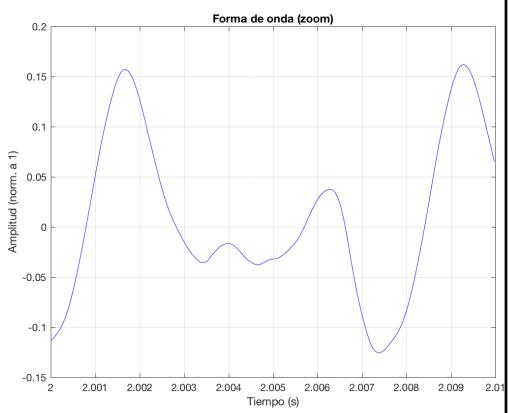
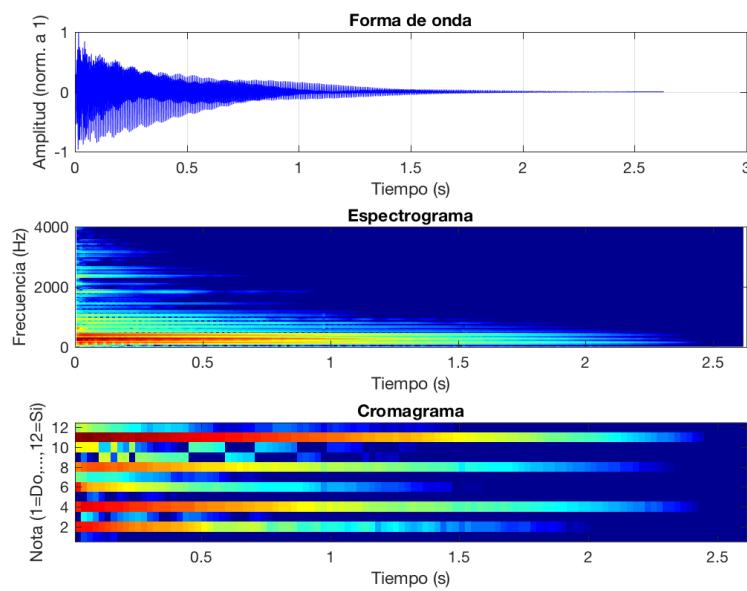
## Ejemplos de Representación de Señales de Audio



Indexación de voz &amp; audio

52

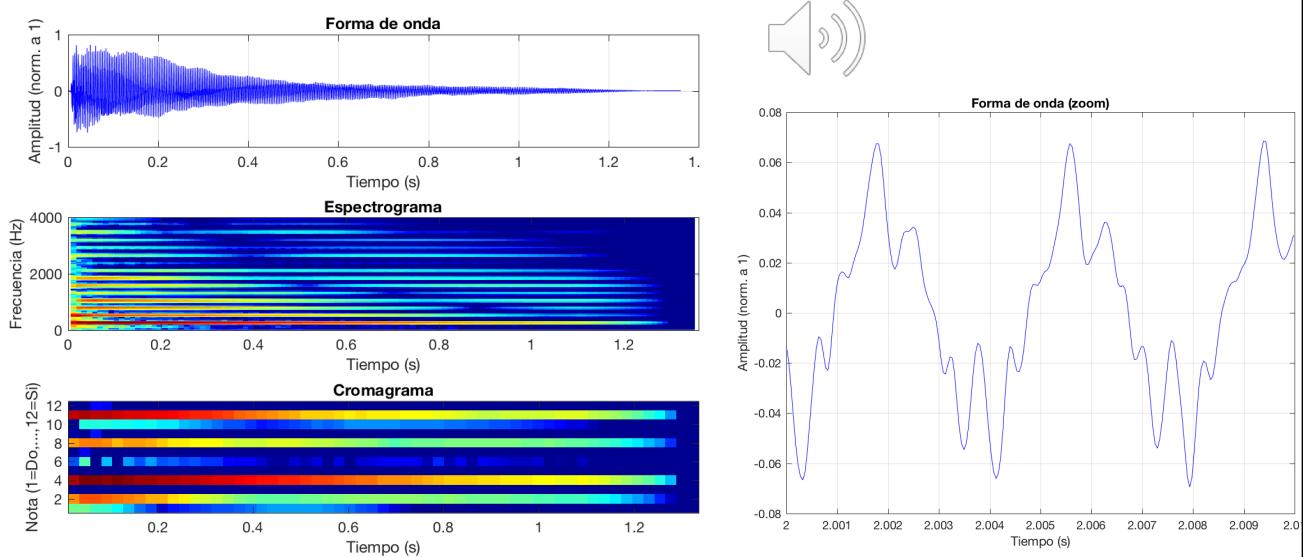
## Ejemplos de Representación de Señales de Audio



Indexación de voz &amp; audio

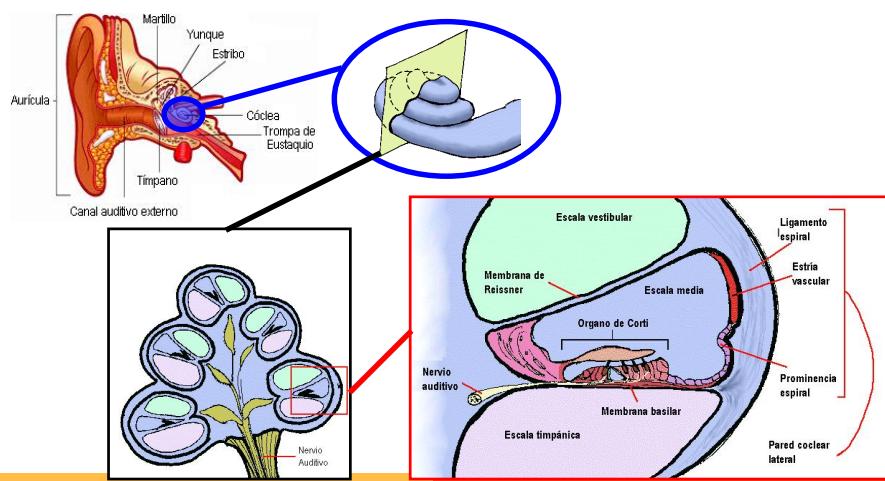
53

## Ejemplos de Representación de Señales de Audio



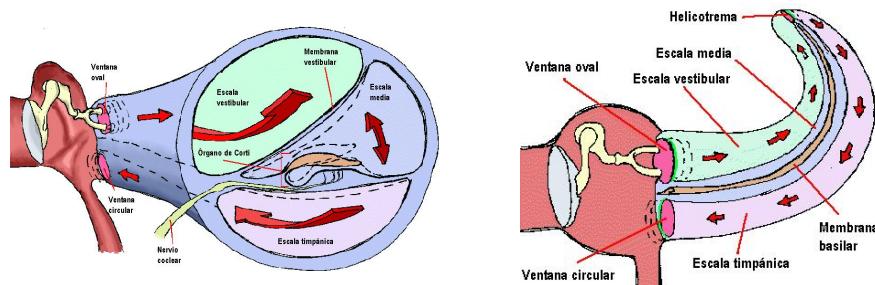
## Introducción al procesamiento de audio y voz

- Las representaciones tiempo-frecuencia suelen estar inspiradas en las características perceptuales del oído humano



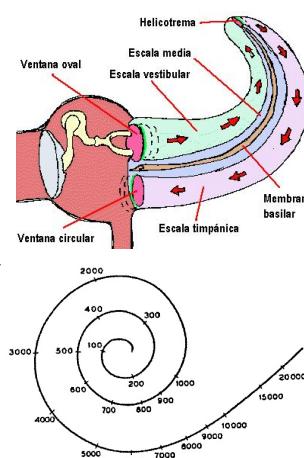
## Introducción al procesamiento de audio y voz

- Y en particular al funcionamiento de la cóclea



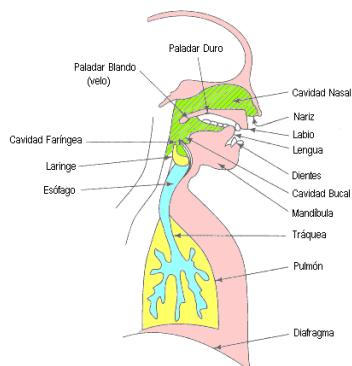
## Introducción al procesamiento de audio y voz

- Funcionamiento de la cóclea:
- La membrana basilar vibra en distintos puntos para distintas frecuencias
  - Se comporta como un analizador frecuencial
  - Máximos de vibración relacionados de forma no lineal con la frecuencia → mayor resolución para bajas frecuencias
- El órgano de Corti transforma vibraciones de la membrana basilar en impulsos nerviosos de forma no lineal



## Producción de voz: la realidad

- La voz es una señal muy particular, porque permite muchas aplicaciones de indexado y búsqueda
- Es muy útil tener un modelo de su funcionamiento, para poder entenderla y reconocerla

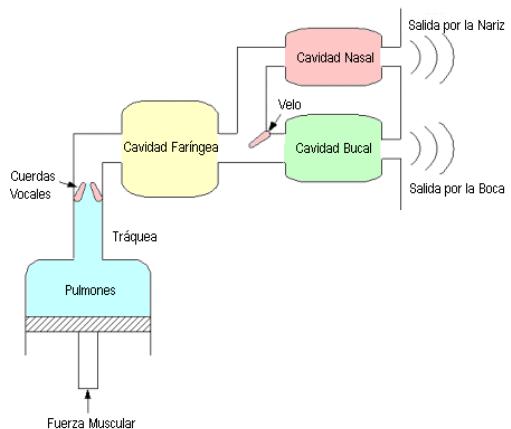
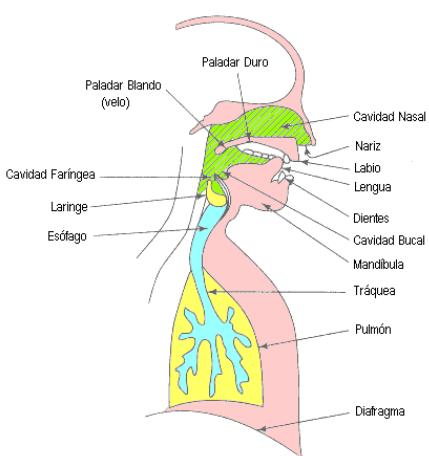


*"Why did Ken set the soggy net on top of his deck"*

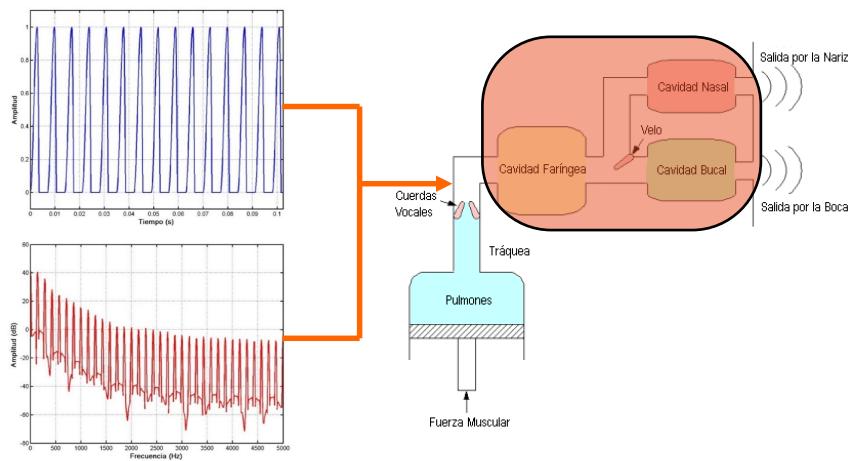
Queen's University/ATR Labs X-Ray Film Database for Speech Research

<https://www.youtube.com/watch?v=DcNMCB-Gsn8>

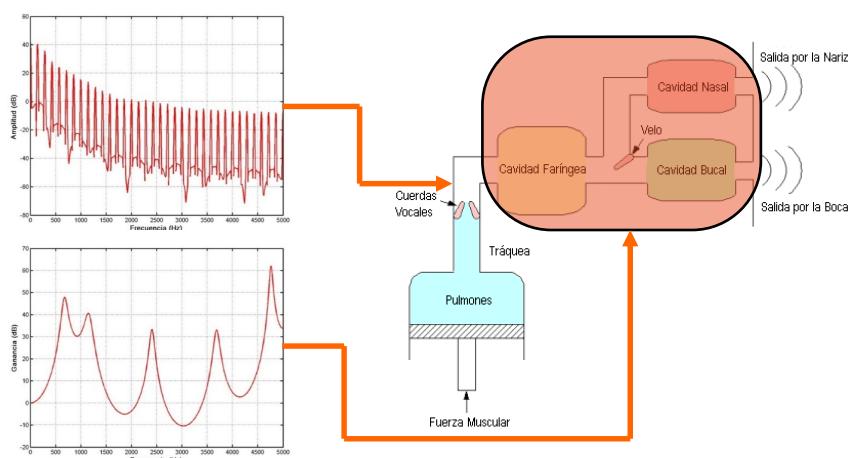
## Producción de voz: funcionamiento simplificado



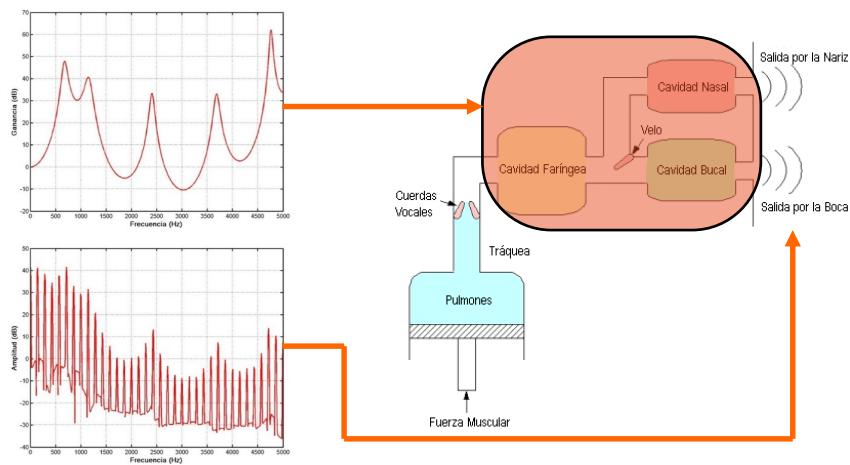
## Producción de voz: simulación (i)



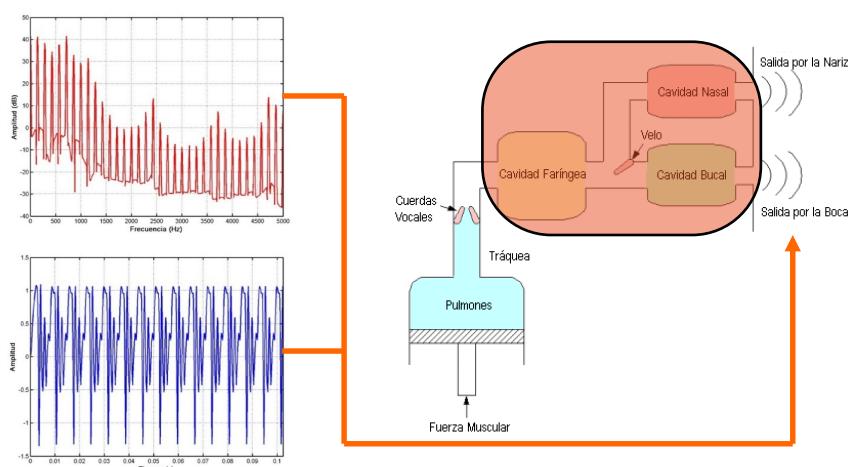
## Producción de voz: simulación (i)



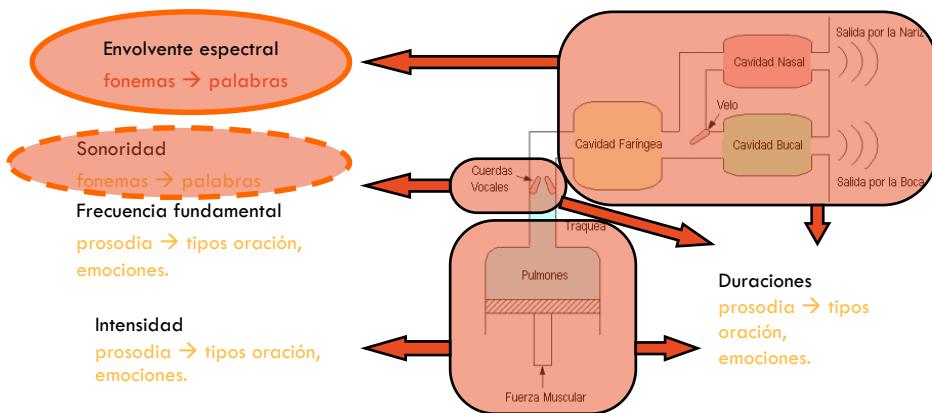
## Producción de voz: simulación (ii)



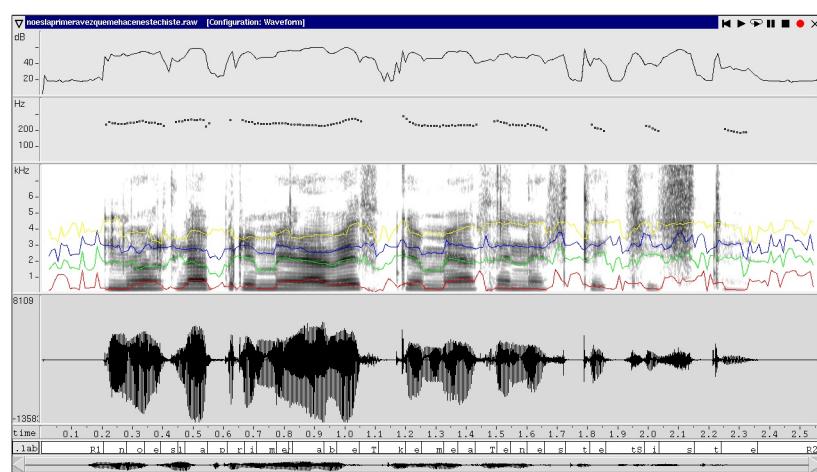
## Producción de voz: simulación (iv)



## Producción de voz: parámetros de la señal de voz afectados

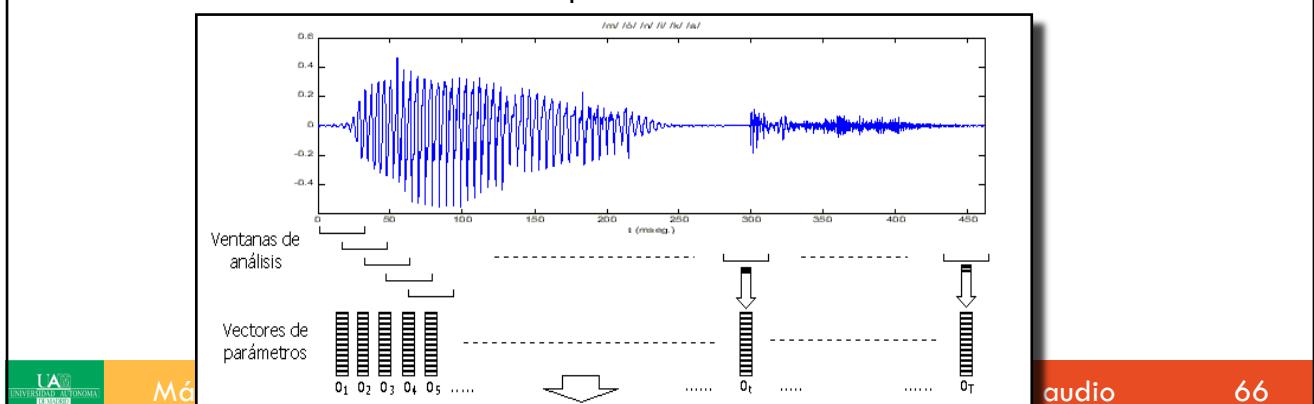


## Producción de voz: el resultado

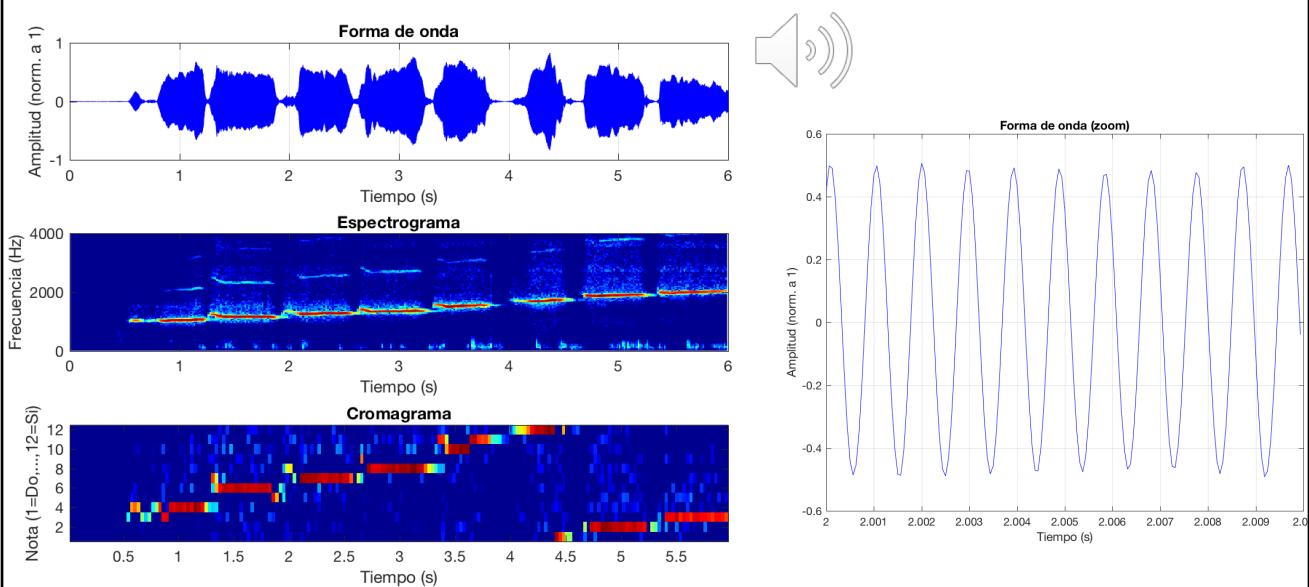


## Parámetros en tratamiento de audio

- Idea: a largo plazo, el audio es muy variable y poco estacionario
- A muy corto plazo, el audio es muy estacionario
- Podemos “trocear” el audio en fragmentos cortos (llamados tramas o ventanas)
- Así, cada trama representará un único tipo de sonido (aproximadamente)
  - Facilita enormemente la clasificación posterior

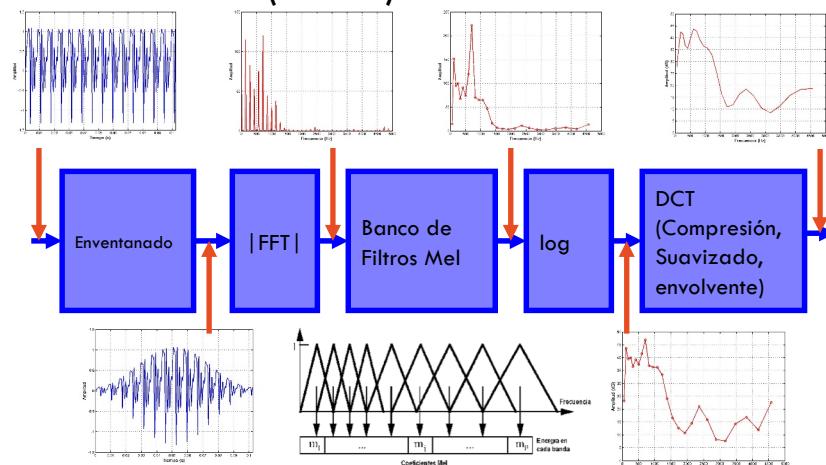


## Ejemplo de Audio a Largo Plazo y a Corto Plazo (estacionario)



## Introducción al procesamiento de audio y voz

- Parámetros más habituales en análisis de voz: Mel-Frequency Cepstral Coefficients (MFCC):

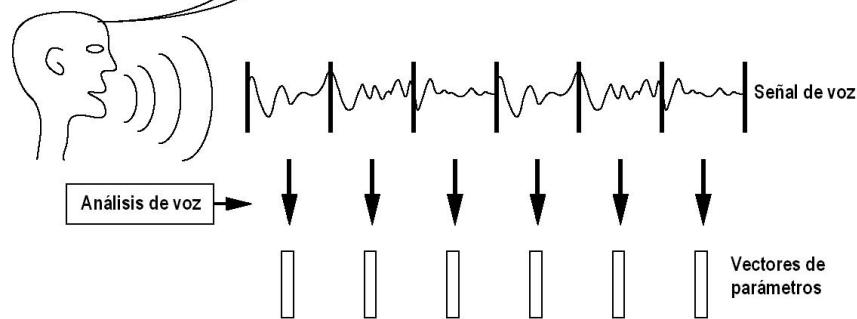


## Introducción al procesamiento de audio y voz

- Con los parámetros MFCC se transforma la voz o el audioen una secuencia de vectores de parámetros sobre la que trabajamos

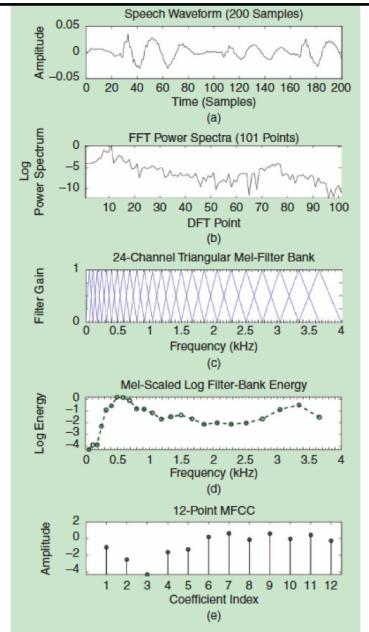
Concepto: Secuencia de símbolos

$S_1 \quad S_2 \quad S_3 \quad \text{etc}$



## Parámetros MFCC

- Capturan información de tipo de sonido
- No capturan información sobre tono (nota musical, etc.)
  - Para eso se usan los cromagramas, por ejemplo
- Sonidos acústicamente diferentes darán lugar a diferentes vectores de parámetros en el espacio de MFCCs
  - Utilizados para distinguir sonidos (fonemas) en voz, locutores, idiomas, tipos de audio...
  - Posibilita una gran variedad de tareas de reconocimiento



[FIG6] Steps in MFCC feature extraction from a speech frame:  
 (a) 200-sample frame representing 25 milliseconds of speech sampled at a rate of 8 kHz, (b) DFT power spectrum showing first 101 points, (c) 24-channel triangular Mel-filter bank, (d) log filter-bank energy outputs from Mel-filter, and (e) 12 static MFCCs obtained by performing DCT on filter-bank energy coefficients and retaining the first 12 values.

## Extracción de Parámetros en Procesamiento de Audio

- Pasos típicos
  - Detección de Actividad de Voz (VAD)
  - Extracción de características
    - Por ejemplo, MFCC, con información dinámica (derivadas, etc.)
  - Compensación de variabilidad a nivel de parámetros
    - Pretenden eliminar variabilidad debido a aspectos diferentes a la información que se pretende extraer y conservar, por ejemplo...
      - Debido al canal de transmisión
      - Debido al ruido de fondo
      - Debido a condiciones acústicas
      - ...

## Referencias de Consulta en Procesamiento de Audio

- **Material no evaluable**
- Fuente: wiki de procesado de Señal de la Universidad de Aalto (FinalIndia):
  - <https://speechprocessingbook.aalto.fi>
- Secciones útiles para ampliar conocimientos en procesamiento de audio y extracción de parámetros
  - 3.1. Short-time analysis of speech and audio signals
  - 3.2. Short-time processing of speech signals
  - 3.3. Waveform
  - 3.4. Windowing
  - 3.5. Signal energy, loudness and decibel
  - 3.6. Spectrogram and the STFT
  - 3.7. Autocorrelation and autocovariance
  - 3.8. The cepstrum, mel-cepstrum and mel-frequency cepstral coefficients (MFCCs)
  - 3.9. Linear prediction
  - 3.10. Fundamental frequency (F0)
  - 3.11. Zero-crossing rate
  - 3.12. Deltas and Delta-deltas



# Indexación de audio: Tema 1: Introducción

< audias >

Audio, Data Intelligence and Speech

<http://audias.ii.uam.es>