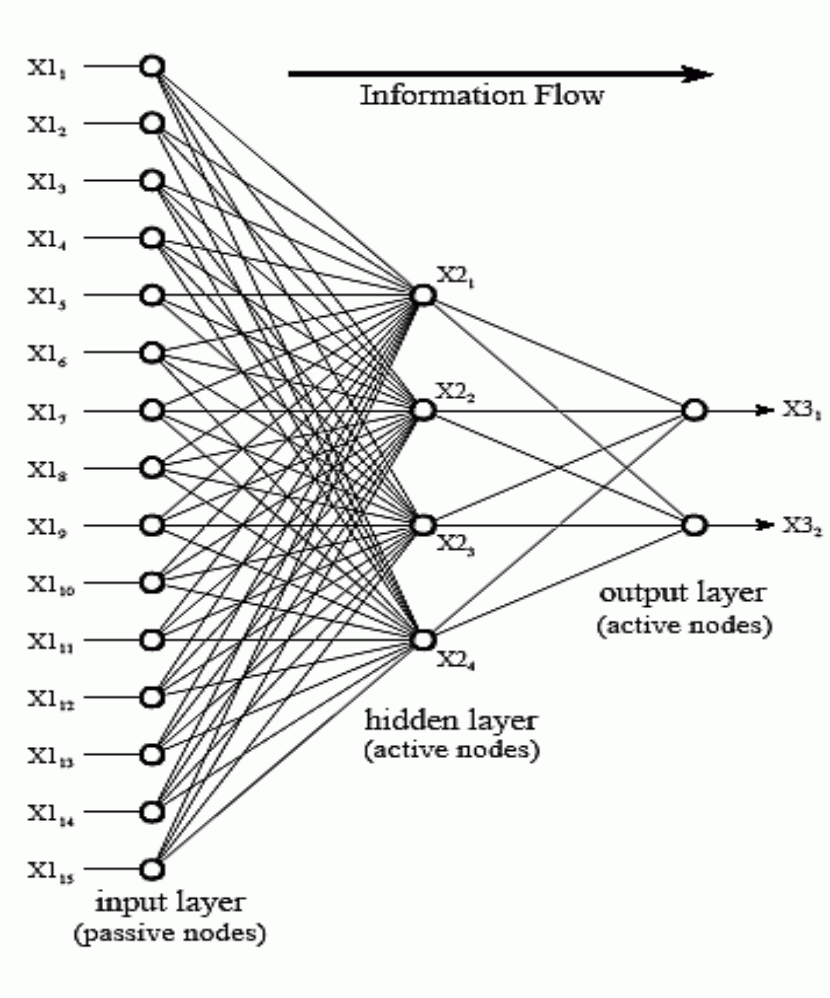
Extracción de Características

Extracción de Características

- Para entrenar modelos de aprendizaje automático (o profundo), no es recomendable alimentar al modelo la señal de audio en crudo:
- Entre más dimensiones de entrada, más complejo debe ser el modelo, y más datos de entrenamiento requiere.



Extracción de Características

- Una “característica” (o feature) no necesariamente tiene que ser algo muy complicado.
- Por ejemplo: el valor promedio de las muestras.
  - Reducción de 48,000 dimensiones a 1.
- Claro, el promedio a lo mejor no es suficiente para ciertas aplicaciones.
- Pero he ahí el punto: **la selección de características tiene que ser apropiada a la aplicación**.

Extracción de Características

- Para llevar a cabo *detección de actividad acústica*: el valor promedio puede ser útil.
- Para llevar a cabo *detección de locutor*: el valor promedio no es suficiente.

Extracción de Características

- La selección de características está relacionada a:

  Si la característica es representativa del elemento a ser reconocido

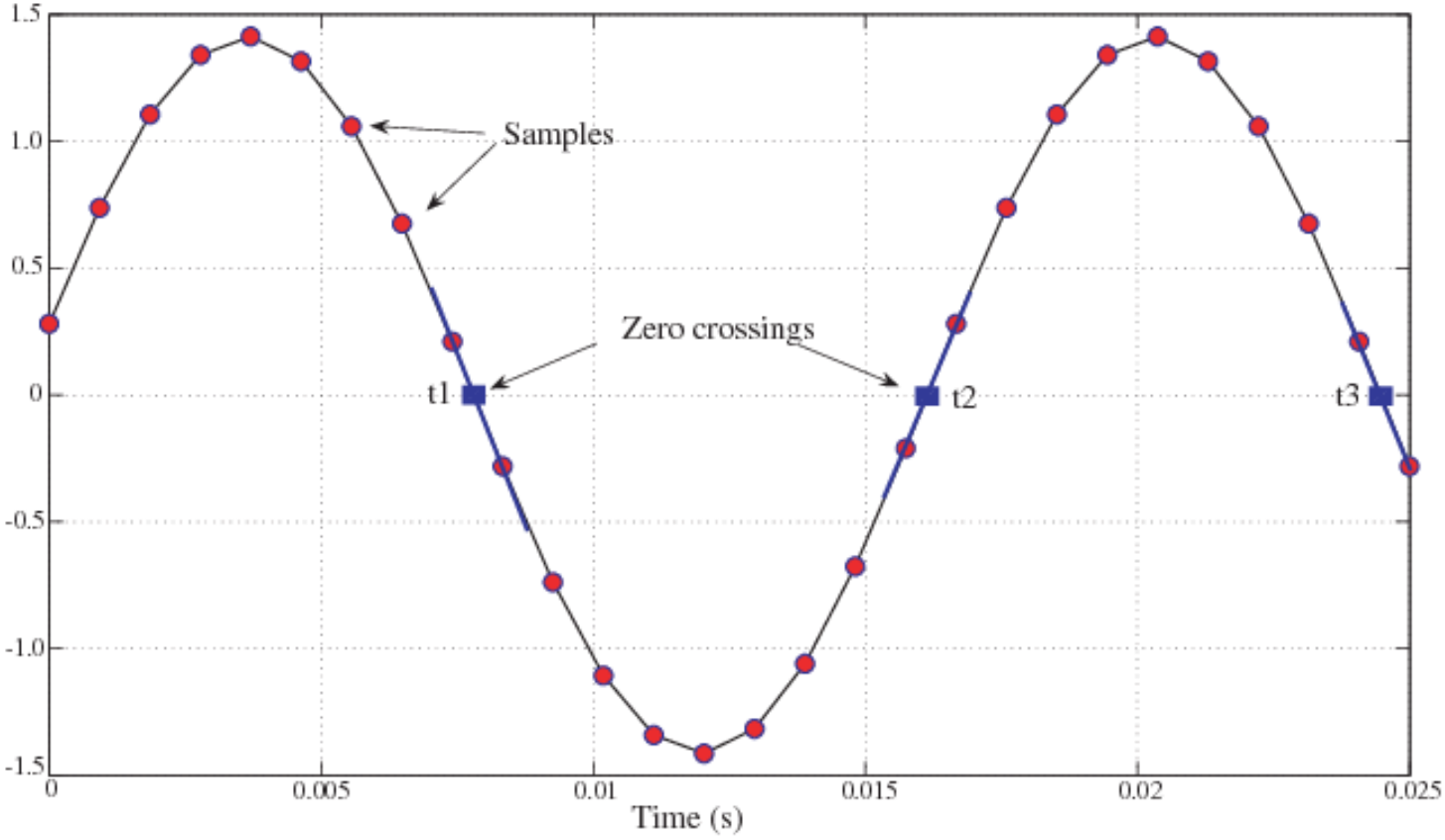
- El promedio de la señal es representativo de la actividad acústica, pero no de quién la esta produciendo.

Extracción de Características

- Hay bastantes.
- Vamos a ver las que en mi opinión son más utilizadas.

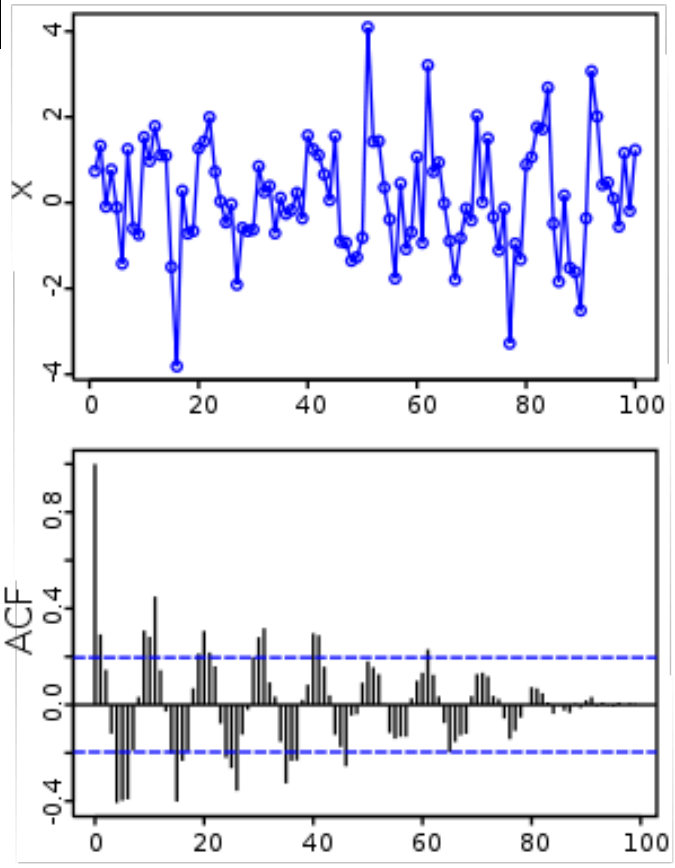
Cruce de cero

- Reduce el número de muestras a utilizar.
- A la vez, proporciona un indicio de frecuencia/ritmo.



Auto-correlación

- Mide qué tanto se parece la señal a versiones de sí misma desfasada.
- Se puede dictaminar la dimensión, acotando la cantidad de desfases medidos.
- Representa indicios de patrones repetidos dentro de la misma señal.

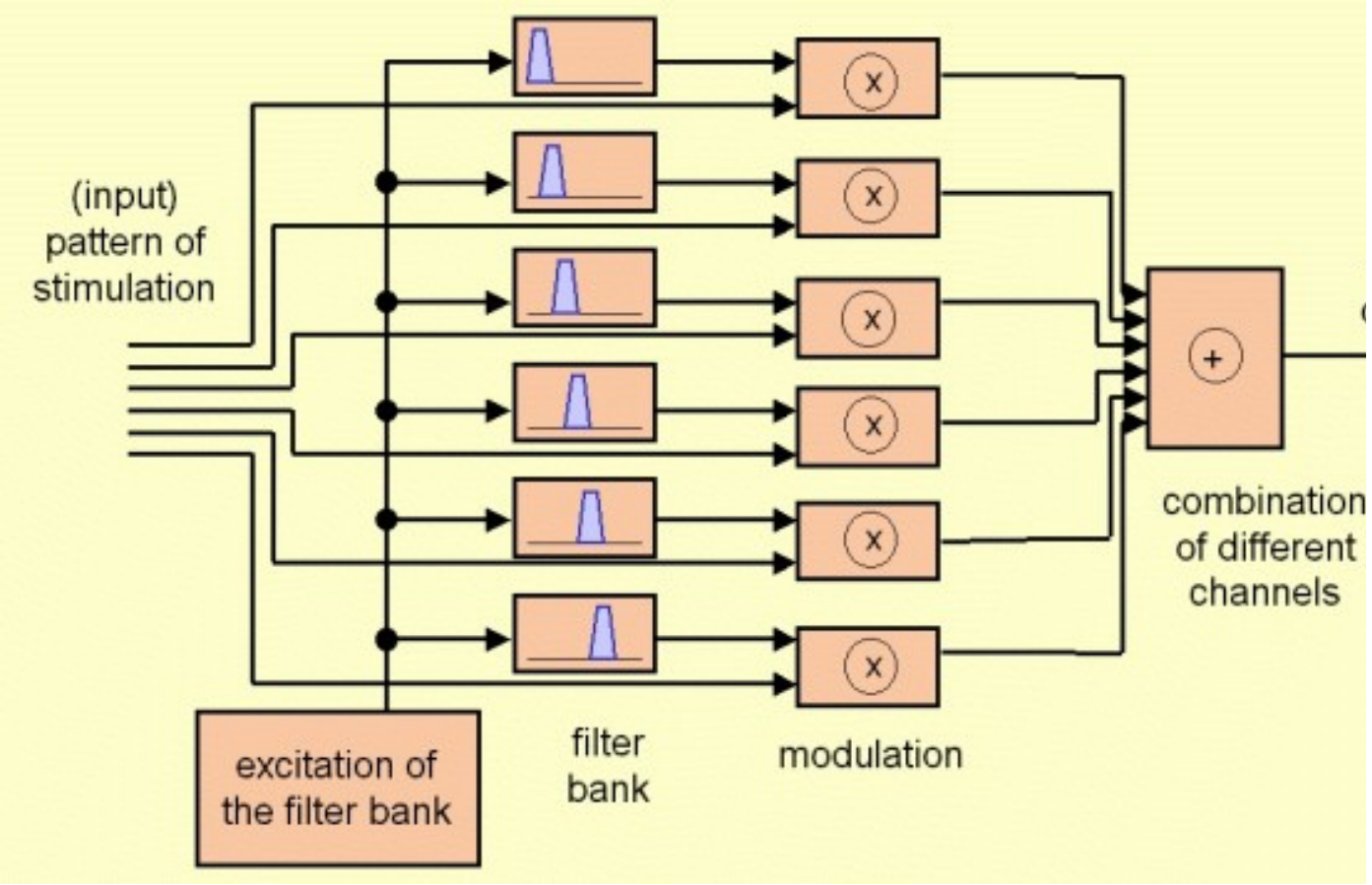


Filtrado Básico

- En el dominio de la frecuencia, se pueden utilizar los valores de solo las frecuencias “útiles”.
  - Voz: 400 a 4000 Hz.

Banco de Filtros

- Generalizando la idea pasada, se pueden “combinar” frecuencias en diferentes rangos.

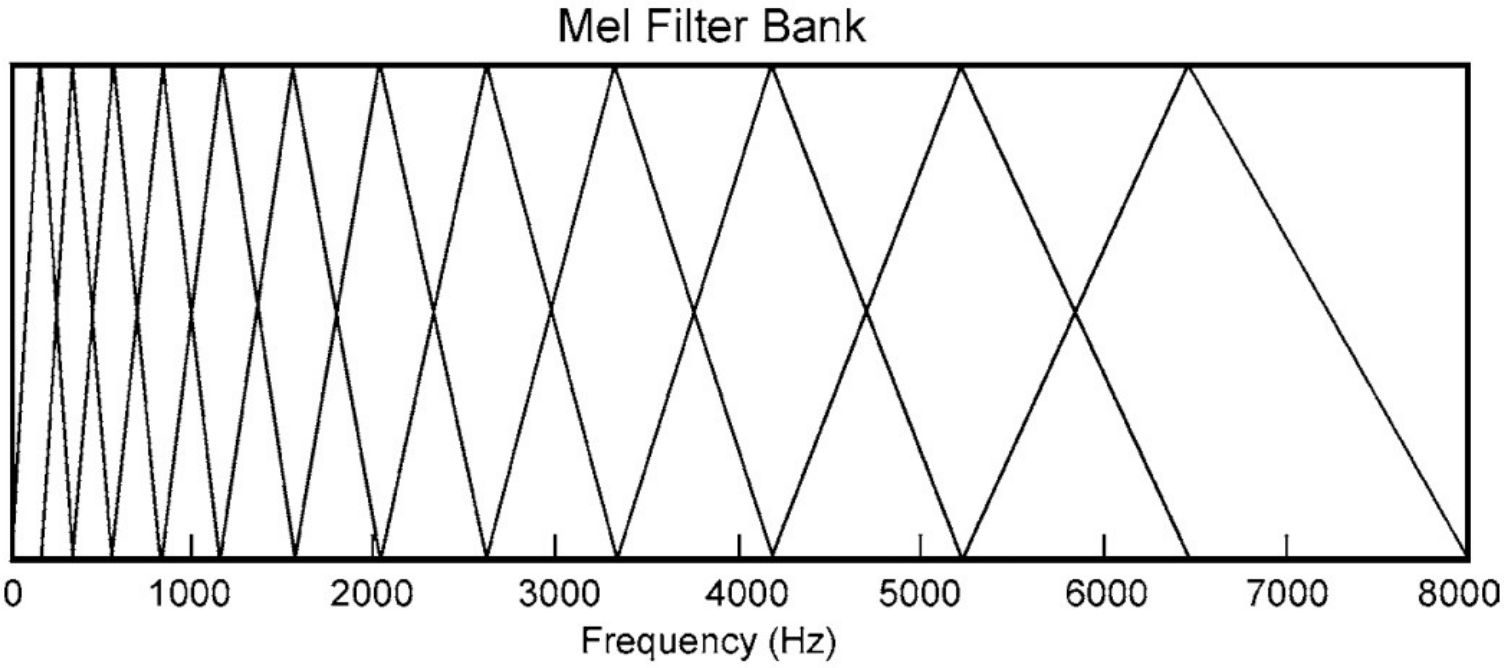


Banco de Filtros

- El número de filtros se convierte en la dimensionalidad de la nueva información a utilizar.
- El problema está en la selección de los filtros:
  - ¿Cual es el rango de cada uno?
  - ¿Cómo están espaciados?
  - ¿Qué forma tienen (cuadrados, triangulares, etc.)?

Banco de Filtros Mel

- Seleccionados para “simular” la manera en la que el oído humano funciona:
  - Altamente discriminatorio en bajas frecuencias.
  - Poco discriminatorio en altas frecuencias.



Banco de Filtros Mel

- Es importante que las características seleccionadas no tengan mucha correlación.
  - Así una dimensión es ortogonal a otra.
  - Y el cambio de una característica no afecta a otras.
- Desafortunadamente los bancos de filtros Mel son bastante correlacionados entre sí.

Coeficientes Ceptrales de Mel (MFCC)

- Para esto, a la salida de los bancos de filtros Mel se le aplica la *Transformada Discreta de Coseno*.
  - La Transformada Discreta de Coseno es parecida a la Transformada de Fourier.
  - En vez de utilizar senos y cosenos, solo utiliza coseno (es más liviana).
- Resultando en una serie de coeficientes ceptrales de Mel (MFCC), que no están correlacionados entre sí, pero siguen representando componentes frecuenciales.
- Además, hay una mayor reducción de la dimensionalidad.
  - Típicamente: 40 filtros de Mel resultan en 12 MFCCs.

Coeficientes Ceptrales de Mel (MFCC)

- Bastante loco:
  - Se le saca una transformada al resultado del filtrado de una señal transformada.

    ¿Transforma-ception?
- Pero, se ha mostrado que una buena selección de MFCCs es representativo de la traquea vocal al momento de vocalizar palabras.
  - Popularmente utilizados para reconocimiento de voz.

Deep-learning Encoders

- Otra forma de extraer características es entrenar un modelo que las extraiga.
- Parece doble trabajo:
  - Entrenar para extraer características, que después se utilizarían para:
  - Entrenar para reconocimiento de patrones

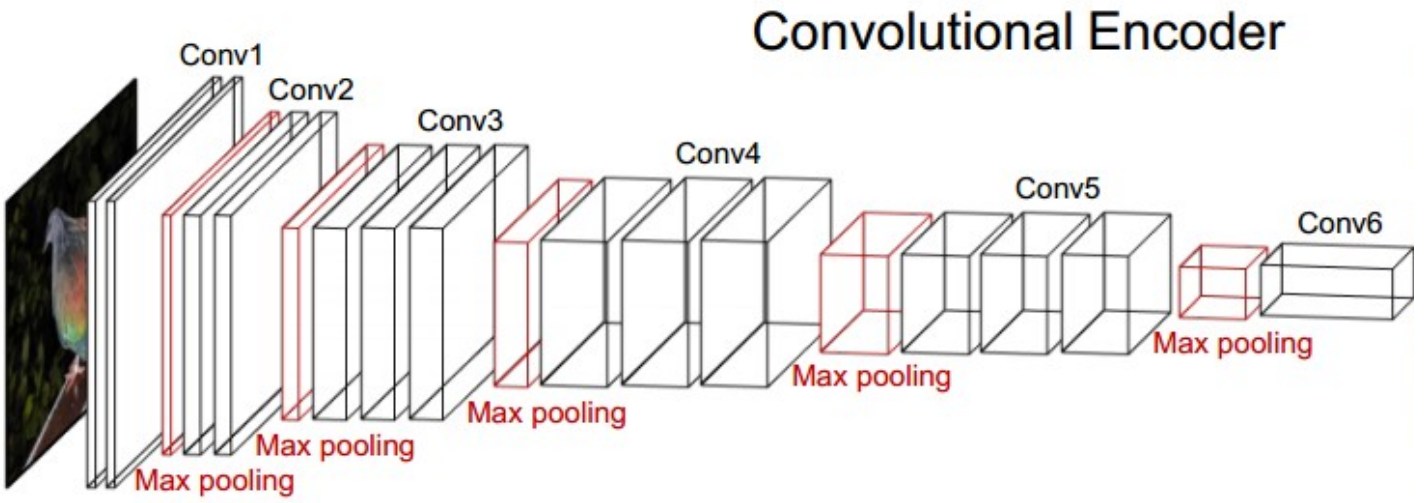
Deep-learning Encoders

- Pero, ultimamente, ha sido una forma muy efectiva de extracción de características representativas a la aplicación.
- En vez de estar “adivinando” cuál característica es mejor, se le da la tarea a la computadora que las diseñe.

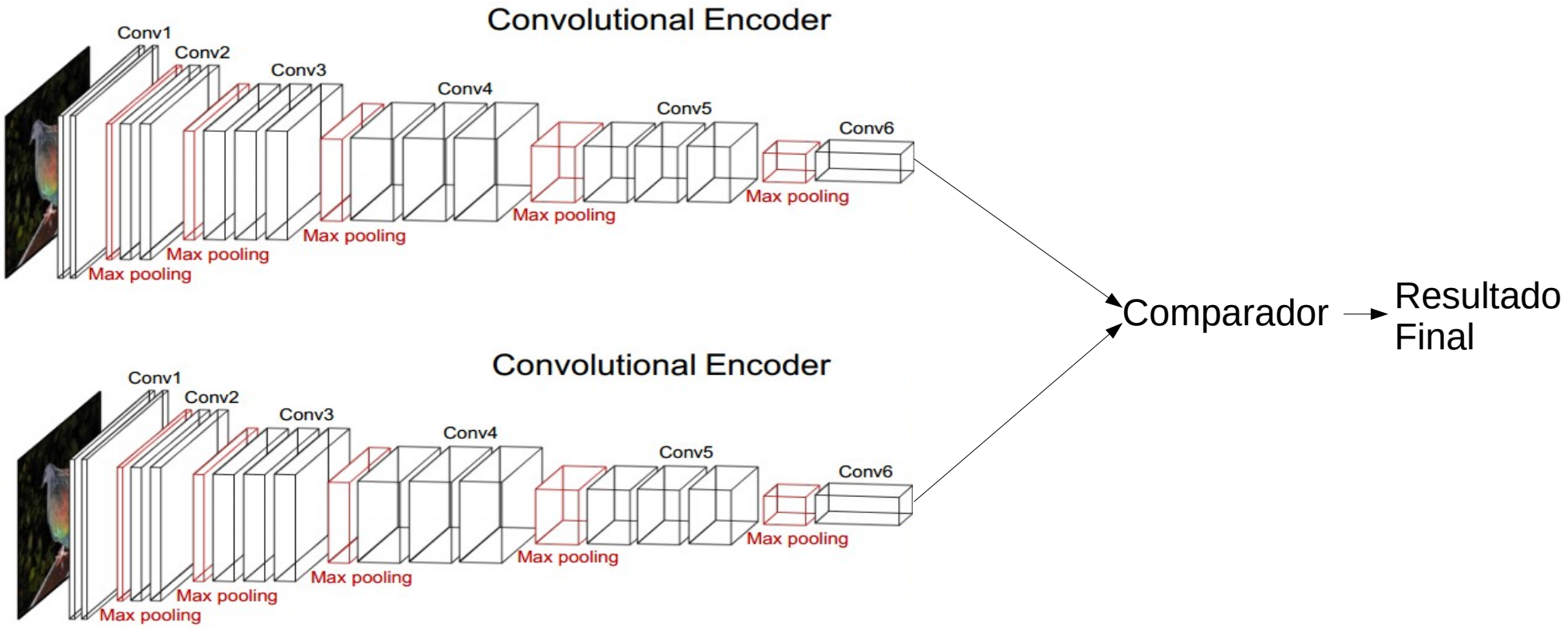
Deep-learning Encoders

- Típicamente, se establece una tarea acotada, pero relacionada a la aplicación.
- *Ejemplo de aplicación*: reconocer si dos señales pertenecen a la misma persona.
- *Ejemplo de tarea acotada*: clasificación de personas conocidas, por medio de voz.

Deep-learning Encoders: Ejemplo Tarea Acotada



Deep-learning Encoders: Ejemplo Aplicación



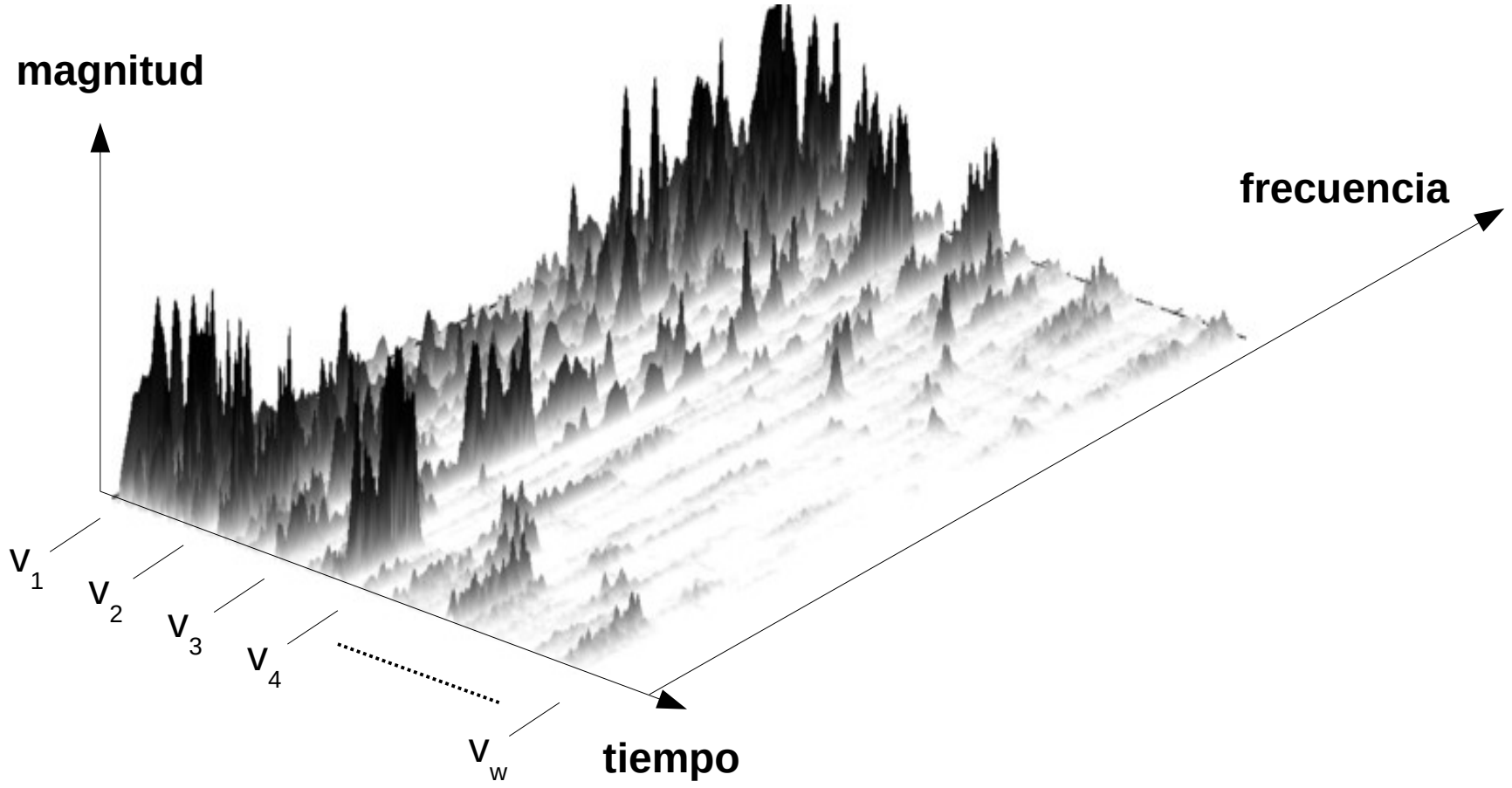
Deep-learning Encoders

- Claro está, queda en misterio lo que significan estas características: ¿qué es lo que están representando de la señal?
  - Se convierten en una caja negra.

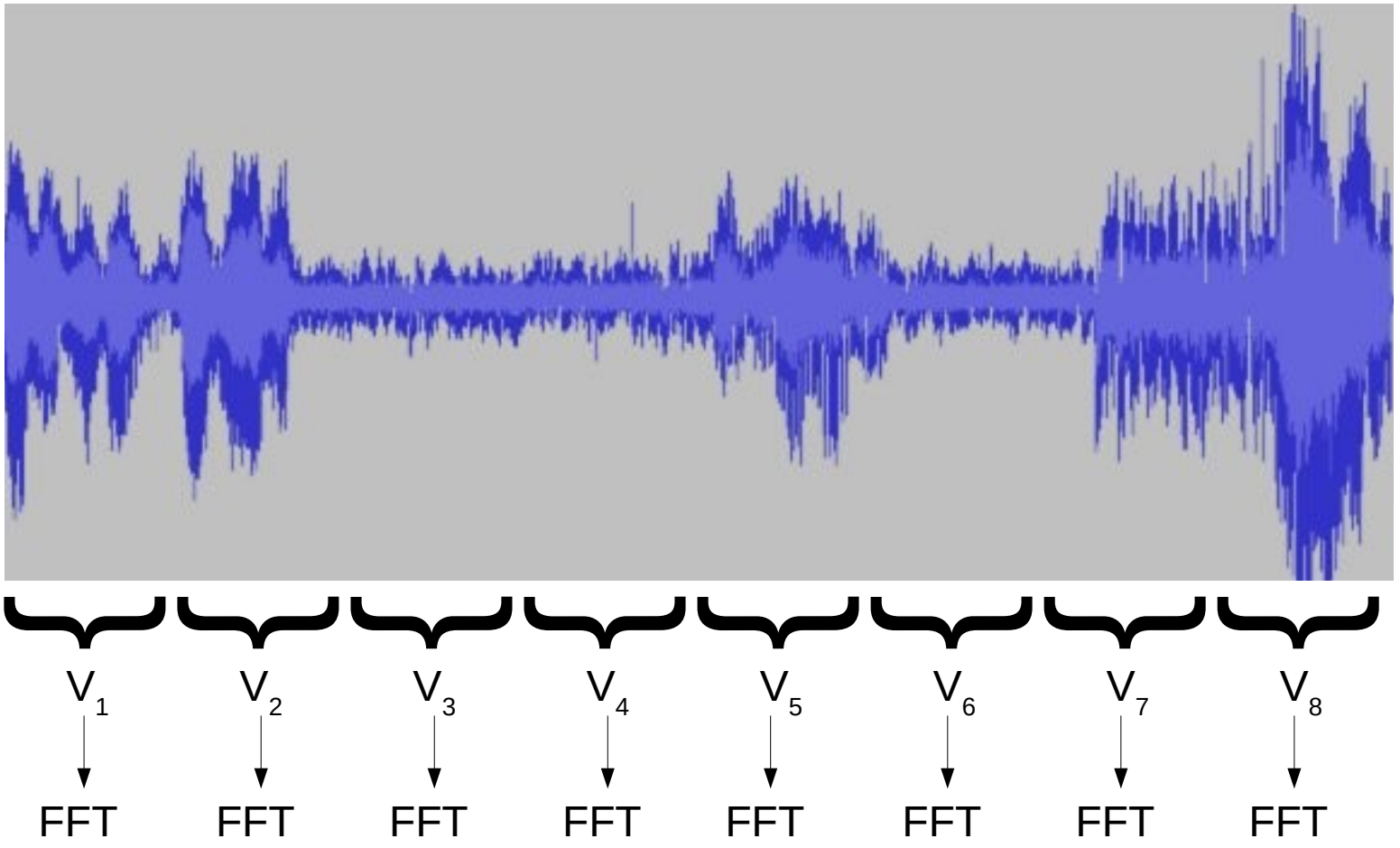
Deep-learning Encoders: Pero...

- “Esos encoders asumen como entrada una imagen, no señales temporales.”
  - Bien observado.
- Afortunadamente, hay una forma muy típica de “imagenizar” una señal temporal:

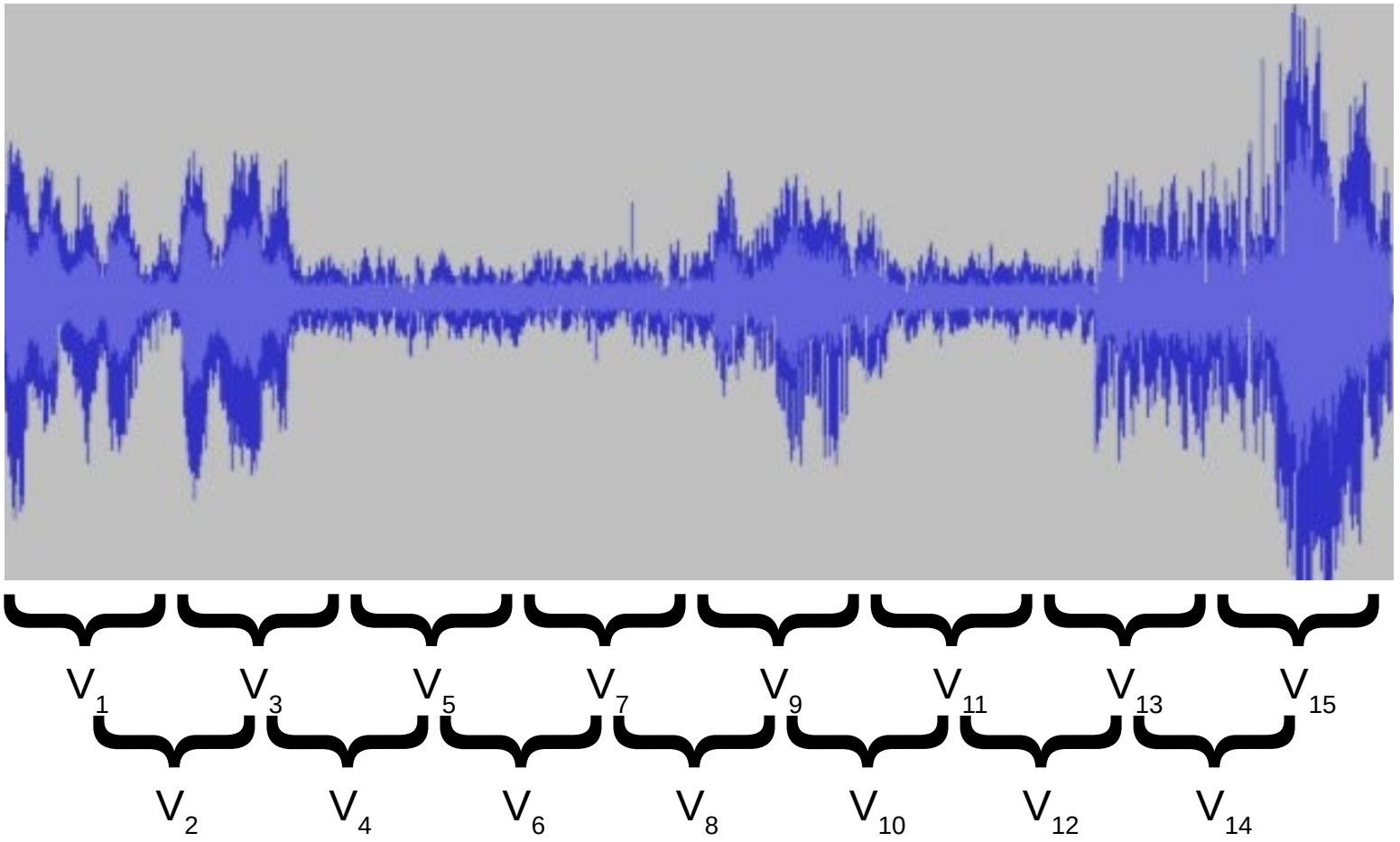
Deep-learning Encoders: Espectrograma



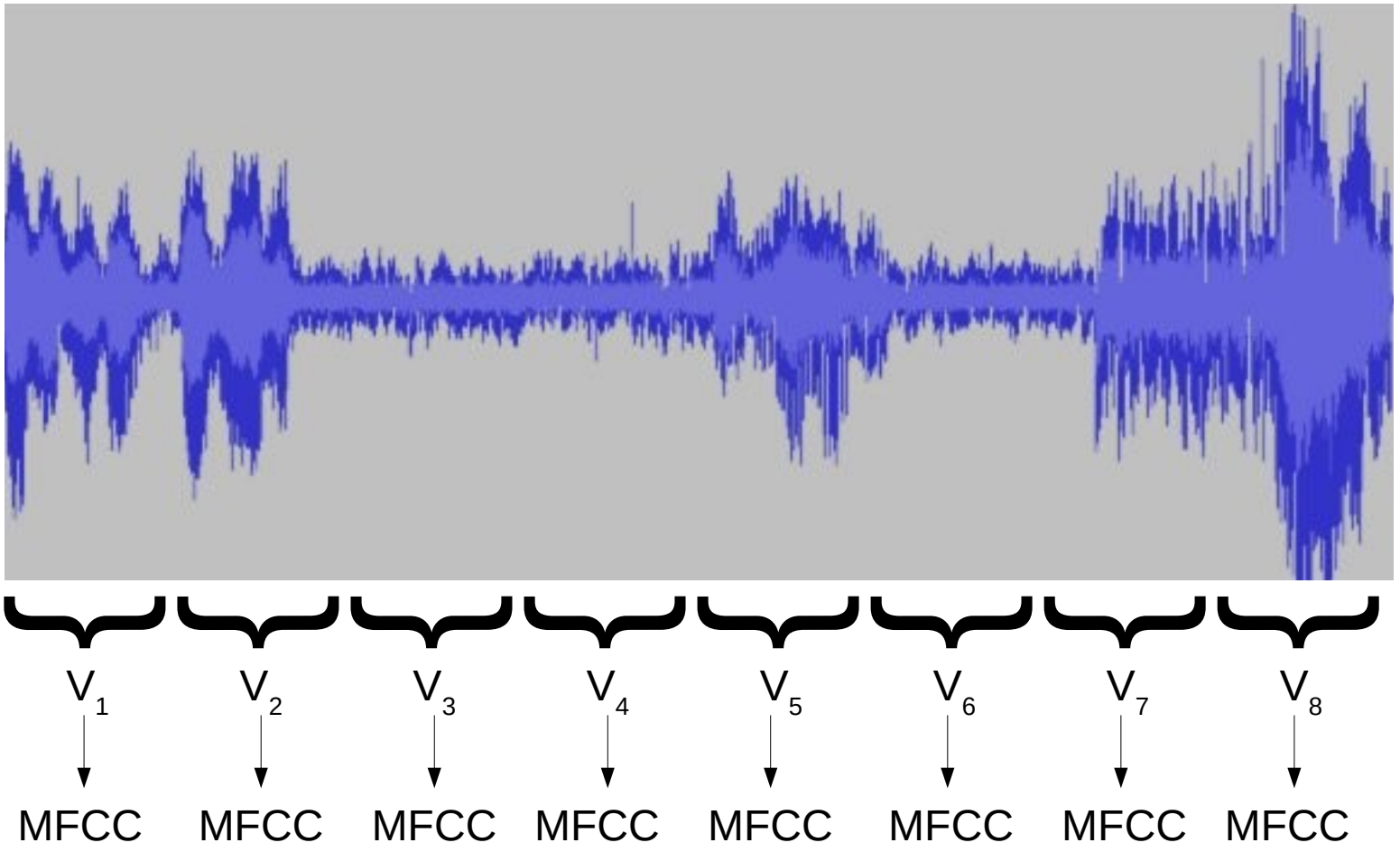
Deep-learning Encoders: Ventaneo



Deep-learning Encoders: Ventaneo para “sacar” más datos



Deep-learning Encoders: Ventaneo con Otras Características

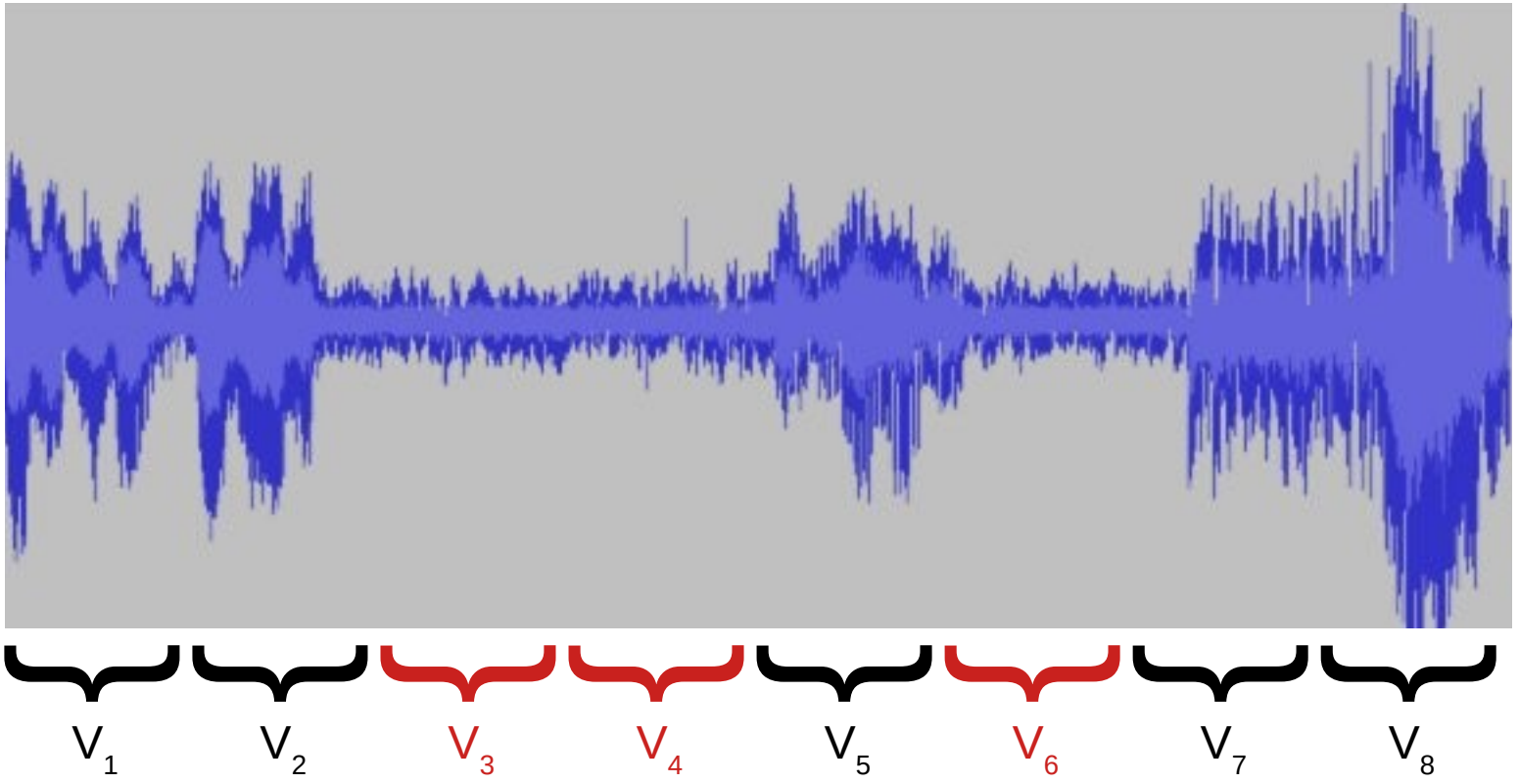


Antes de terminar...

Selección de Ventanas

- Es posible que los datos capturados en una ventana tengan poca energía o que no haya actividad.

Selección de Ventanas



Selección de Ventanas

- Si utilizan estas ventanas “silenciosas”, corren el riesgo de confundir a su sistema.
  - La misma información (el silencio) corresponderá a más de una clase.
- Para esto, es importante ignorar estas ventanas “silenciosas” y sólo utilizar ventanas “activas”.

Detección de Actividad de “Voz” (VAD)

- Ya lo platicamos en la sesión pasada, pero quiero recalcar algo aquí.
- Hay muchas técnicas que se pueden utilizar, pero es importante no caer en ese hoyo negro.
  - Diseñar sistemas de VAD es un área de investigación por si misma.
- Por lo que, sí, el paso del VAD es importante, pero no se “claven” en que el VAD sea perfecto.

Y...

Más Características a Extraer

- Sharma, Garima, Kartikeyan Umapathy, and Sridhar Krishnan. "**Trends in audio signal feature extraction methods**." Applied Acoustics 158 (2020): 107020.
  - Completa revisión de características para audio, y que pueden ser utilizadas en cualquier señal temporal.
  - Pseudo-código en Matlab de cada una.
- Hay una copia del artículo en la página del curso.