DIA 1 | 2:30 hrs

Procesamiento de Senales de Audio (I)

Sonido digital, muestreo, Nyquist, FFT, espectrogramas, filtros Mel y MFCCs

¿Que es el sonido?

El sonido es una vibracion

que viaja por el aire como una ola en el agua. Cuando hablas:

1. Tus cuerdas vocales vibran
2. Crean ondas de presion en el aire
3. Las ondas llegan al microfono
4. El microfono las convierte en numeros

La computadora solo entiende numeros.

Senal de audio digital

Conceptos clave de una onda de sonido:

Amplitud

Que tan 'fuerte' suena (altura de la onda). Enojo = amplitud alta.

Frecuencia

Que tan 'agudo' o 'grave' suena (ondas por segundo, en Hz).

Duracion

Cuanto tiempo dura el sonido. Nosotros usamos 2 segundos.

Representacion digital: frecuencia de muestreo

¿Que es la frecuencia de muestreo?

Es cuantas veces por segundo 'fotografiamos' la onda de sonido.

Cada 'fotografia' captura la altura de la onda en ese instante. A mas fotos por segundo, mas fiel es la copia digital.

Nosotros usamos 16,000 Hz, es decir, tomamos 16,000 muestras cada segundo.

Analogia: flipbook

Un flipbook con 3 dibujos por segundo se ve cortado. Con 24 dibujos/seg se ve fluido. Con el audio es igual: pocas muestras = audio cortado, muchas muestras = audio fiel.

Frecuencias de muestreo comunes

8,000 Hz	Telefono fijo	Baja (solo voz basica)
16,000 Hz	Nuestro proyecto	Suficiente para voz
22,050 Hz	Radio AM	Buena para musica simple
44,100 Hz	CD de musica	Alta fidelidad
48,000 Hz	Video/cine	Profesional
96,000 Hz	Estudio grabacion	Ultra alta fidelidad

¿Por que 16 kHz? Porque la voz humana tiene frecuencias entre 85 Hz (grave) y 8,000 Hz (aguda). No necesitamos capturar mas.

Teorema de Nyquist: ¿por que exactamente el doble?

$$fs >= 2 \times fmax$$

Para capturar fielmente una frecuencia fmax, necesitas muestrear al menos al doble (2 × fmax). Si muestreos mas lento, pierdes informacion (aliasing).

Ejemplo practico

Voz humana: fmax ≈ 8,000 Hz
fs >= 2 × 8,000 = 16,000 Hz

Si usaramos fs = 8,000 Hz:
- Solo captariamos hasta 4,000 Hz
- Perderiamos las consonantes (s, f, t)
- El audio sonaria 'apagado'

fs = 16,000 Hz es perfecto para voz.

Analogia: el cine

Una rueda gira a 10 vueltas/seg.

Si filmas a 20+ fps: se ve girando normal.
Si filmas a 10 fps: parece DETENIDA.
Si filmas a 8 fps: parece girar al REVES.

Eso es 'aliasing' — capturar mal por muestrear muy lento. Nyquist dice: minimo 2× la frecuencia mas rapida.

Por eso el cine usa 24 fps (para movimientos de hasta 12 ciclos/seg).

¿Que es la frecuencia? Graves vs agudos

Frecuencia = ondas por segundo (Hz)

1 Hz = 1 onda por segundo
100 Hz = 100 ondas/seg (tono grave)
1,000 Hz = 1,000 ondas/seg (tono medio)
8,000 Hz = 8,000 ondas/seg (tono agudo)

El oido humano escucha de 20 Hz a 20,000 Hz.
La voz habla entre 85 Hz y 8,000 Hz.

¿Por que importa para las emociones?

Enojo	Tono ALTO, variable, energia en agudos
Tristeza	Tono BAJO, monotono, poca energia
Alegria	Tono alto, melodico, rango amplio
Miedo	Tono alto, temblor, irregular

La escala Mel: como el oido percibe las frecuencias

El oido humano NO percibe las frecuencias de forma lineal. La diferencia entre 100 Hz y 200 Hz se escucha ENORME, pero entre 8,000 Hz y 8,100 Hz apenas se nota.

La escala Mel compensa esto: da MAS resolucion a frecuencias bajas (donde el oido es sensible) y MENOS a frecuencias altas. Por eso usamos filtros 'Mel' y no filtros lineales.

Transformada de Fourier Rapida (FFT)

¿Que hace la FFT?

Cualquier sonido complejo es la SUMA de muchas ondas simples (senos y cosenos).

La FFT descompone un sonido en sus frecuencias individuales, como un prisma separa la luz blanca en colores.

Entrada: senal en el tiempo (amplitud vs tiempo)
Salida: espectro (amplitud vs frecuencia)

Ejemplo con musica

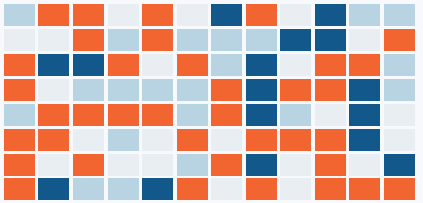
Imagina que escuchas una banda:
🎸 Guitarra: frecuencias medias
🥁 Bateria: frecuencias bajas
🎤 Voz: frecuencias altas

Tu oido los escucha todos JUNTOS.
La FFT los SEPARA para que puedas analizar cada instrumento por separado.

Con la voz, la FFT nos dice que frecuencias estan activas en cada instante.

De la FFT al espectrograma

Si aplicamos la FFT a muchas ventanitas consecutivas del audio (cada 32ms), obtenemos un ESPECTROGRAMA. Es una 'radiografia' del sonido donde: Eje X = tiempo (cada ventana) | Eje Y = frecuencia (graves abajo, agudos arriba) | Color = intensidad (oscuro = silencio, brillante = energia).



Paso 1: Pre-enfasis

$$y[n] = x[n] - 0.97 \cdot x[n-1]$$

Resta el 97% de la muestra anterior a la actual. Esto amplifica las frecuencias ALTAS y atenua las BAJAS.

¿Por que?

La voz tiene MUCHA mas energia en frecuencias bajas que en altas. Sin pre-enfasis, los graves 'esconderian' a los agudos.

Es como ajustar el ecualizador del carro para subir los agudos: no cambia la cancion, pero ahora escuchas las consonantes mas claro.

Las consonantes (s, f, t, k) tienen energia en frecuencias altas y son CLAVE para distinguir palabras y emociones.

Ejemplo numerico

```
Muestras: [100, 102, 98, 105]

y[0] = 100 (primera, se queda)
y[1] = 102 - 0.97×100 = 5.0
y[2] = 98 - 0.97×102 = -0.94
y[3] = 105 - 0.97×98 = 9.94

Resultado: [100, 5.0, -0.94, 9.94]

Los cambios rapidos (agudos) se amplifican.
Los valores 'planos' (graves) casi
desaparecen.
```

Paso 2: Ventaneo con Hamming

¿Que es el ventaneo?

Dividir el audio en pedacitos (ventanas) para analizar cada uno por separado.

Cada ventana = 512 muestras = 32 ms
Avanzamos 256 muestras (50% superposicion)

¿Por que superposicion? Para no perder informacion en los bordes de cada ventana.

¿Por que Hamming?

Si cortamos el audio 'a lo bruto', los bordes de cada ventana tienen saltos abruptos que crean frecuencias falsas en la FFT (fuga espectral).

La ventana de Hamming suaviza los bordes: multiplica cada muestra por un valor entre 0 y 1, donde el centro vale ~1 y los bordes valen ~0.08.

Analogía: es como hacer fade-in/fade-out en cada pedacito de audio.

Numeros del ventaneo

Audio total	32,000 muestras (2 seg × 16,000 Hz)
Ventana	512 muestras = 32 ms
Hop (salto)	256 muestras = 16 ms (50% superposicion)
Total ventanas	(32,000 − 512) / 256 + 1 = 124 frames

Paso 3: FFT de cada ventana

Aplicar la FFT a cada una de las 124 ventanas

Cada ventana tiene 512 muestras en el TIEMPO. La FFT la convierte a 257 valores en FRECUENCIA (de 0 Hz a 8,000 Hz).
Resultado: matriz de 257 × 124 (frecuencias × ventanas). Esto ya es un espectrograma en crudo.

¿Por que 257? La FFT de N puntos produce N/2 + 1 valores unicos → 512/2 + 1 = 257

Cada valor de la FFT nos dice:

1. A QUE frecuencia corresponde
   (bin 0 = 0 Hz, bin 256 = 8,000 Hz)

2. CUANTA energia hay a esa frecuencia
   (valor alto = mucha energia)

El enojo se caracteriza por mucha energia en frecuencias medias-altas (500-4000 Hz).

Analogia: el prisma

Un prisma toma luz blanca (mezcla de colores) y la separa en sus colores individuales.

La FFT toma un sonido complejo (mezcla de frecuencias) y lo separa en sus frecuencias individuales.

Luz blanca → rojo, naranja, amarillo...
Voz → 200 Hz, 500 Hz, 1000 Hz...

Así podemos 'ver' que frecuencias usa una persona enojada vs una persona tranquila.

Paso 4: Filtros Mel triangulares

¿Que son los filtros Mel?

Son 20 filtros con forma de triangulo que agrupan las 257 frecuencias de la FFT en 20 bandas.

Las bandas NO son iguales:
- Bandas bajas (graves): angostas → mas detalle
- Bandas altas (agudos): anchas → menos detalle

Esto imita al oido humano que distingue mejor los graves que los agudos.

¿Por que triangulos?

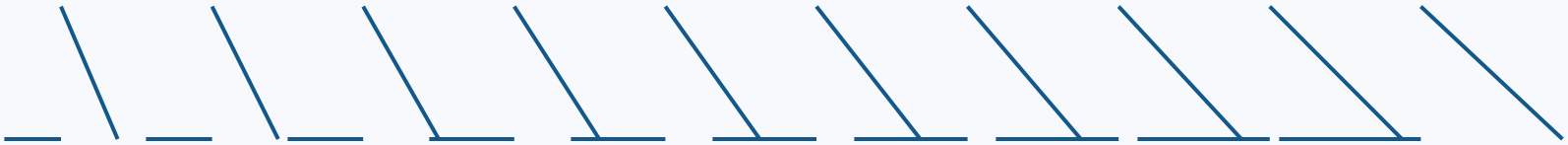
Cada triangulo tiene:
- Base: rango de frecuencias que cubre
- Pico: frecuencia central (peso maximo)
- Lados: transicion gradual

El valor de cada banda es la suma ponderada de las energias de las frecuencias cubiertas por el triangulo.

Resultado: de 257 valores de FFT pasamos a solo 20 valores Mel. Compresion 13x.

$$mel(f) = 2595 \times \log_{10}(1 + f/700)$$

Banco de 20 filtros Mel (forma visual)



Paso 5: Logaritmo + DCT

Logaritmo: comprimir rango dinamico

Los valores de energia Mel varian ENORMEMENTE: de 0.001 a 1,000,000.

El logaritmo comprime este rango:
log(0.001) = -6.9
log(1) = 0
log(1,000,000) = 13.8

Ahora los valores van de -7 a 14 en vez de 0 a 1,000,000. Mucho mas manejable para la red neuronal.

Ademas, el oido percibe el volumen de forma logaritmica (los decibeles son logaritmicos).

DCT: compresion final

La Transformada Discreta del Coseno (DCT) es como una segunda FFT pero sobre los coeficientes Mel.

¿Por que? Los 20 valores Mel estan correlacionados (bandas vecinas son similares). La DCT decorrelaciona estos valores.

Resultado: 20 coeficientes INDEPENDIENTES que capturan la 'forma' del espectro.

Analogía: es como comprimir una foto JPEG. Los primeros coeficientes capturan lo mas importante, los ultimos son detalles finos.

Resultado final: cada ventana de 512 muestras se resume en 20 coeficientes MFCC.

Con 124 ventanas × 20 coeficientes = matriz de 20 × 124. Luego recortamos al centro: 20 × 63 = 1,260 numeros. De 32,000 muestras crudas a 1,260 numeros esenciales — una compresion de 25x sin perder la informacion clave.

Pipeline MFCC completo: resumen visual

	Paso	Descripcion	Metafora	Forma
1	**Pre-enfasis**	Resalta agudos: y = x − 0.97·x_ant	*Subir agudos en ecualizador*	`32,000`
2	**Ventaneo Hamming**	Dividir en 124 ventanas de 512 muestras	*Cortar salchicha en rodajas*	`512 × 124`
3	**FFT**	Cada ventana: tiempo ⟶ frecuencia	*Prisma separa colores*	`257 × 124`
4	**Filtros Mel (20)**	Agrupar 257 frecuencias en 20 bandas	*Ecualizador de 20 bandas*	`20 × 124`
5	**Logaritmo**	Comprimir rango dinamico enorme	*Volumen en decibeles*	`20 × 124`
6	**DCT**	Decorrelacionar los 20 coeficientes	*Comprimir foto a JPEG*	`20 × 124`
7	**Recorte central**	Tomar 63 frames del centro (la voz)	*Quitar silencios inicial/final*	`20 × 63`

Preprocesamiento: limpiar antes de analizar

Eliminacion de ruido

El mundo real tiene ruido: ventiladores, trafico, electrodomesticos. Usamos filtros digitales para quedarnos solo con la voz humana (85-8000 Hz). El pre-enfasis tambien elimina ruido DC (offset electrico del microfono). Sin filtrar, la red confunde ruido con enojo.

Deteccion de actividad de voz (VAD)

No todo el audio contiene voz. Medimos la energia por ventana del espectrograma: si hay un salto >= 5 dB entre ventanas consecutivas, AHI empieza la voz. Si no hay voz, el sistema no gasta energia haciendo prediccion. Ahorra bateria y evita falsos positivos.

Balanceo de clases

En el dataset ESD hay 4 emociones 'Other' (Happy, Neutral, Sad, Surprise) vs 1 'Angry'. Ratio: 4:1. Si entrenamos asi, el modelo predice siempre 'Other' y acierta 80% sin aprender NADA. Solucion: tomar la misma cantidad de cada clase con random sampling.

Ruido blanco para robustez

Agregamos ruido gaussiano MUY suave (desviacion estandar / 50) a cada audio de entrenamiento. ¿Para que? Asi el modelo no memoriza audios perfectos de laboratorio y funciona mejor cuando hay ruido de fondo real. Es como entrenar con condiciones 'imperfectas'.