

# Detección de Emociones

Kun Han, Dong Yu, Ivan Tashev - 2014

# Reconocimiento de Emociones

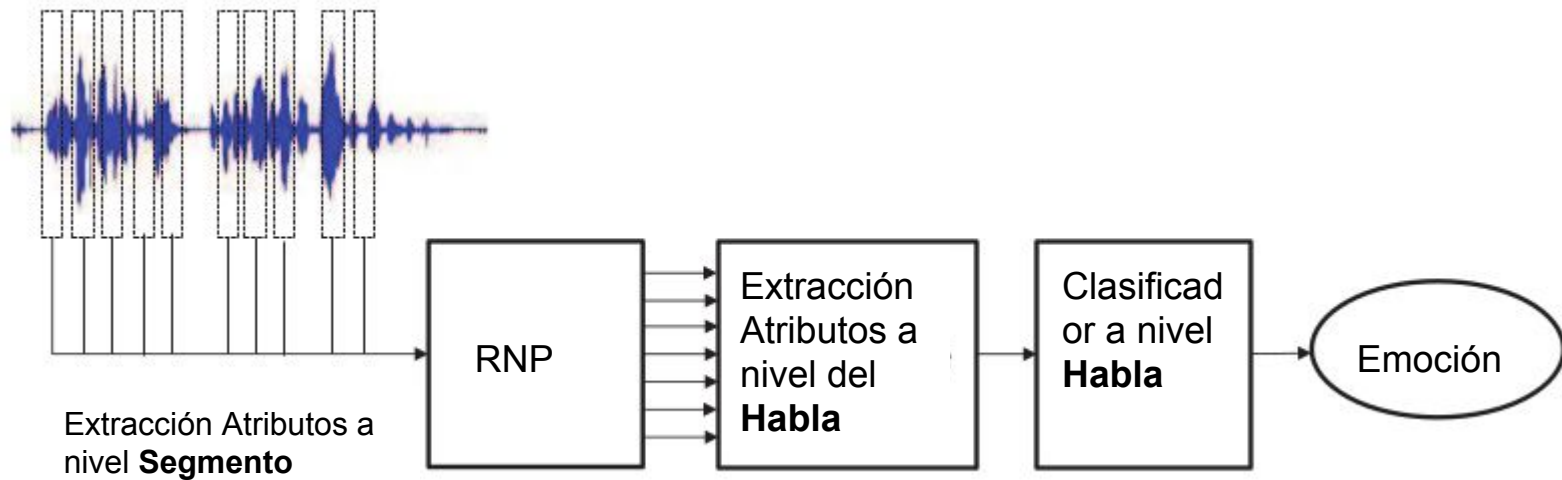
¿Qué características o atributos del habla nos ayudan?


Previamente eran seleccionados a mano.

Redes Neuronales Profundas

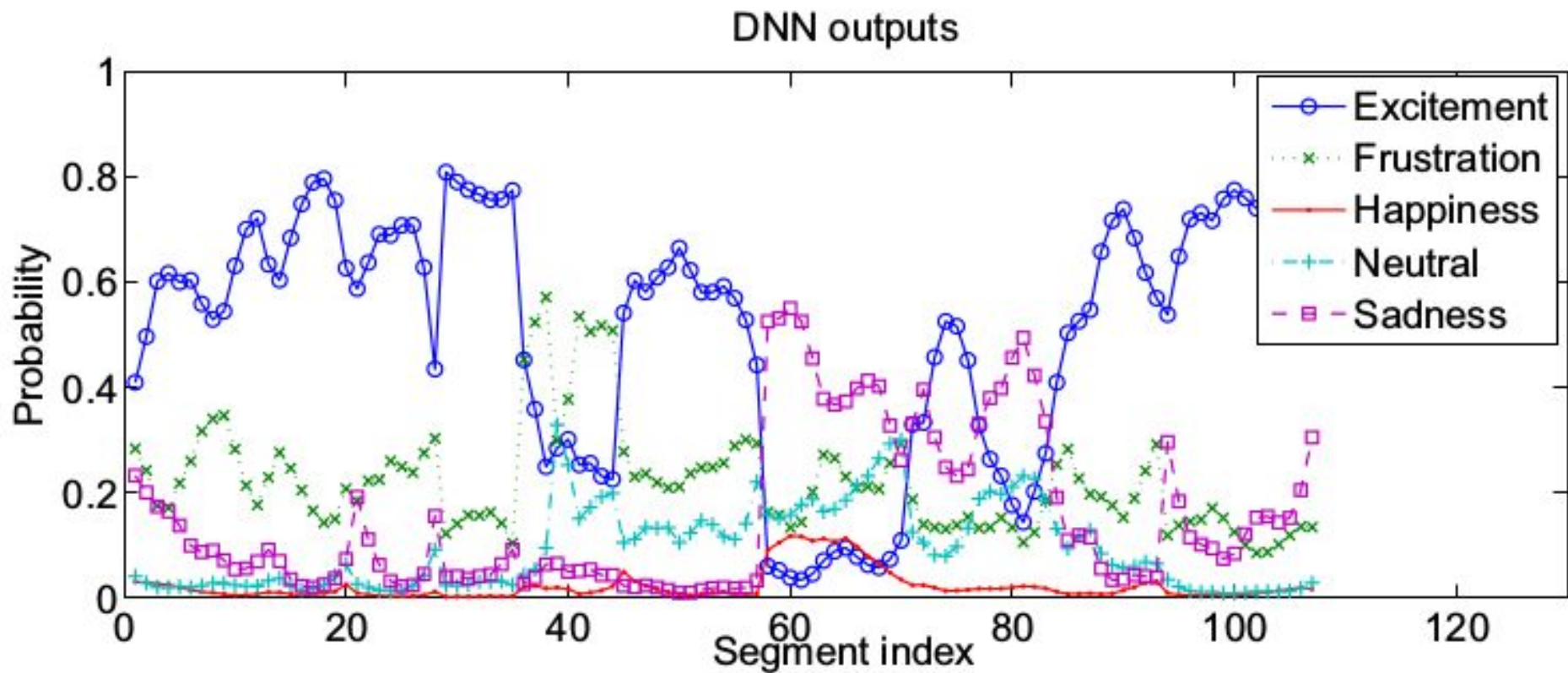


# Esquema



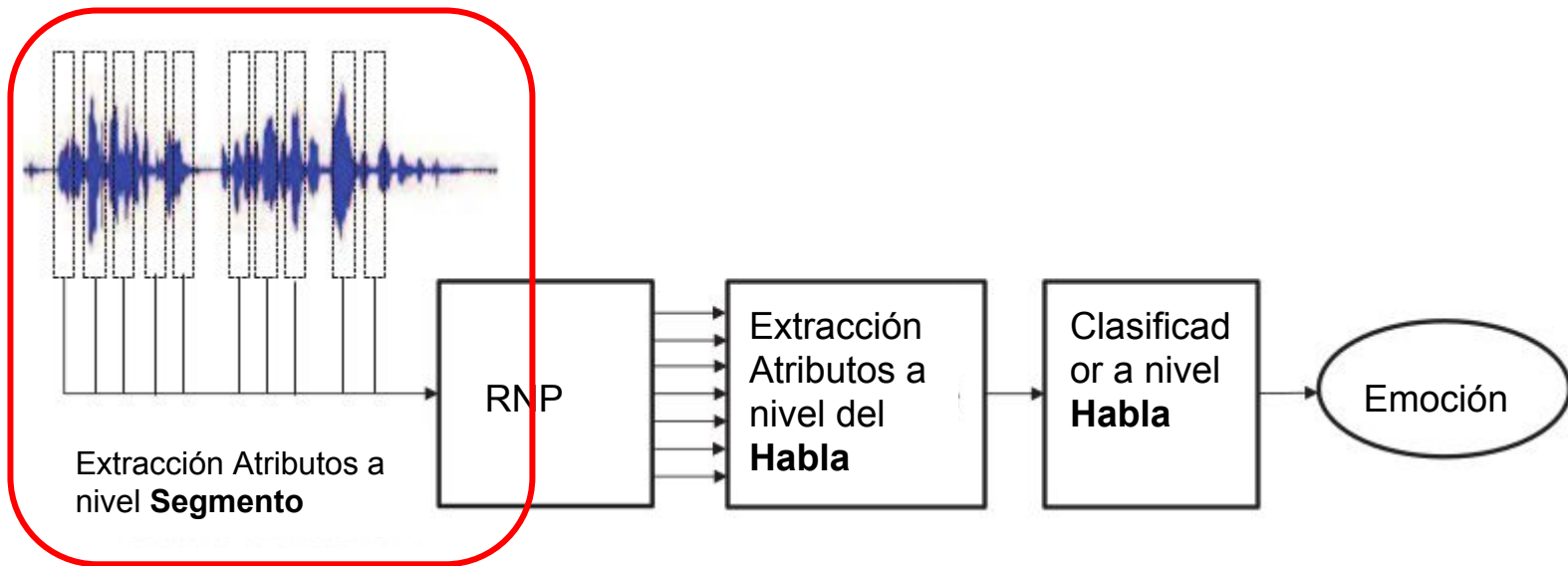


Emoción de los  
segmentos  
determina la  
Emoción del audio  
completo



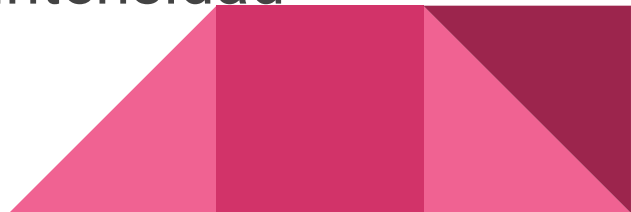
Salida de la RNP de un audio

# Esquema

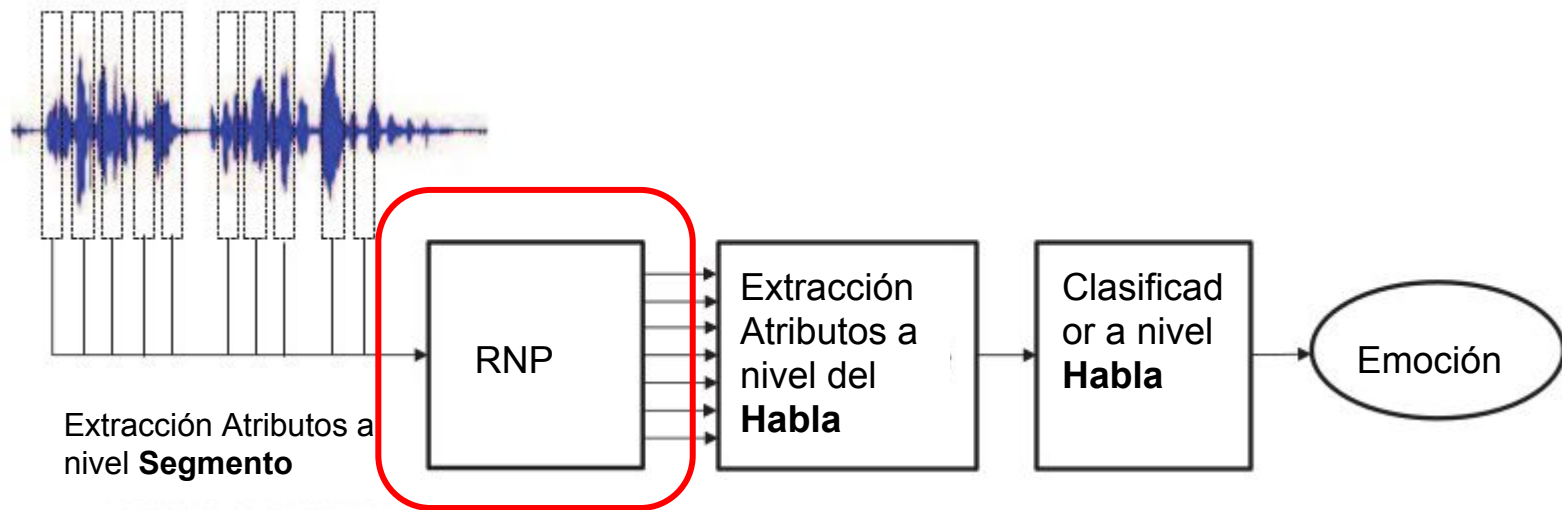


# Extracción de Atributos a nivel **Segmento**

1. Separan la señal en segmentos de 25 ms, cada 10 ms.
2. Generan un vector de atributos formado por
  - a. Atributos MFCC
  - b. Atributos tonales
  - c. Diferenciales entre ellos a través del tiempo
3. Se descartan los segmentos con poca intensidad



# Esquema





# RNP y su entrenamiento

## Entrenamiento

- Conjunto de vectores de atributos.
- X segmentos / vectores por cada audio en el conjunto de entrenamiento.
- La salida esperada es la emoción etiquetada para el audio entero.

## Una vez entrenada

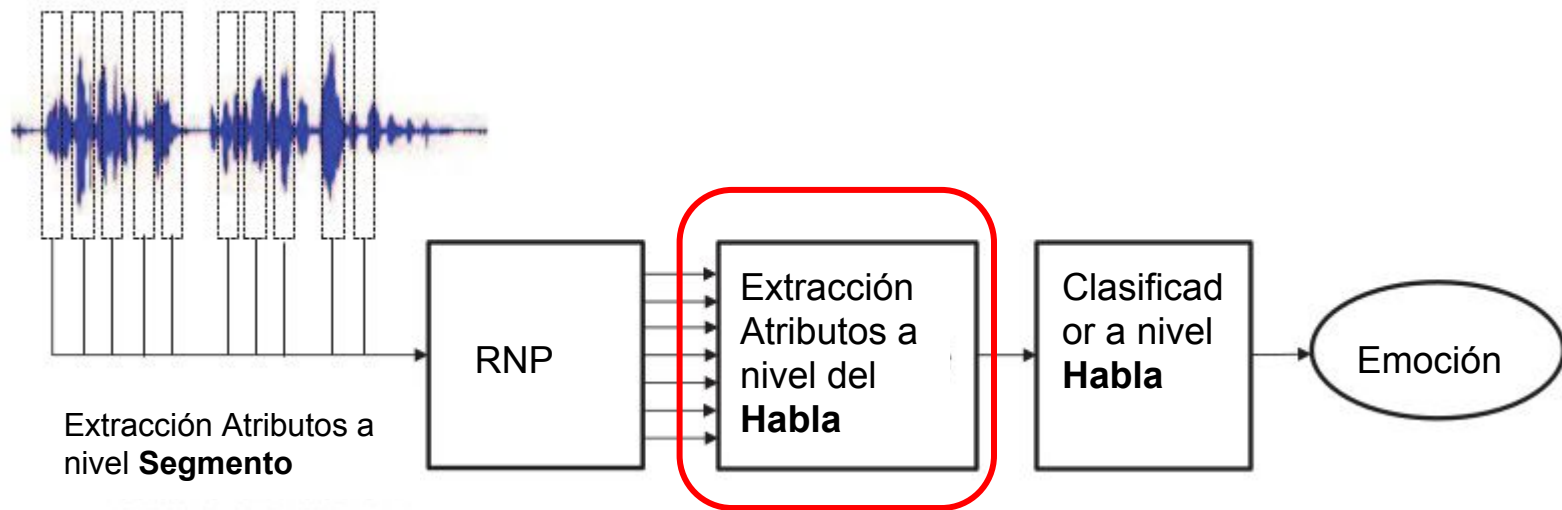
- Entrada: Vector de Atributos de un segmento de una señal de audio
- Salida: Distribución de probabilidades de las Emociones
  - $T = [P(E_1), \dots P(E_k)]$

# Estructura de la RNP

- Capas:
  - 1 de entrada. Con 750 unidades, correspondientes al tamaño del vector.
  - 3 ocultas, con 256 unidades
  - 1 de salida, con K unidades



# Esquema

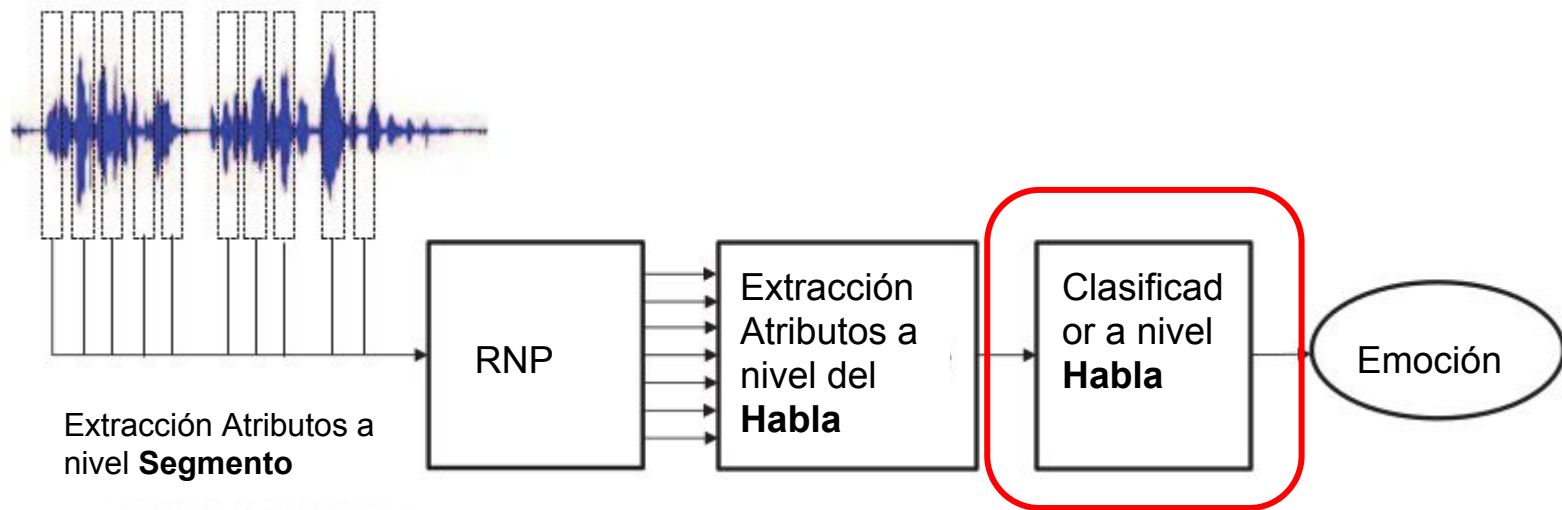


# Extracción de Atributos a Nivel **Habla/Audio**

- Se computan en base a estadísticas sobre probabilidades obtenidas de los segmentos
- Denotando  $P_s(E_k)$  = “prob. Emoción  $k$  para segmento  $s$ ”
  - $F_1^k = \mathbf{Max}_{s \in U} P_s(E_k)$
  - $F_2^k = \mathbf{Min}_{s \in U} P_s(E_k)$
  - $F_3^k = \mathbf{Avg}_{s \in U} P_s(E_k)$
  - $F_4^k = \%_{s \in U} P_s(E_k) > \theta$

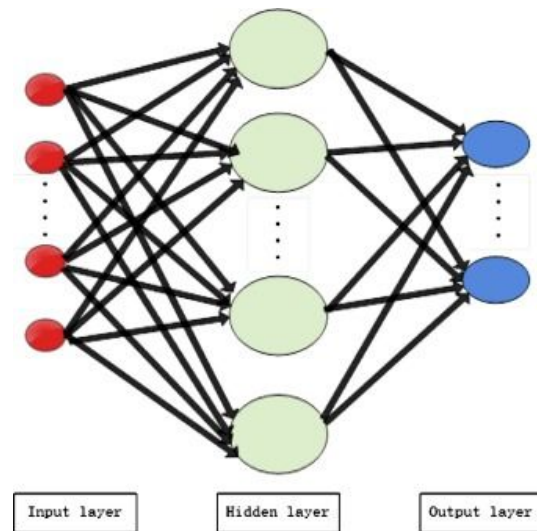


# Esquema



# Extreme Learning Machine (ELM)

- Red neuronal con 1 capa oculta
- Capa oculta es mucho mayor a la dimensión de la entrada
- Entrenamiento no convencional
  - Capa 1: Pesos fijos
  - Capa 2: Entrenamiento rápido (Matrices)
- Buenos resultados para conjuntos de entrenamiento pequeños



# Clasificación a nivel Habla

## Entrenamiento ELM

- Conjunto de vectores  $F$ , de atributos estadísticos.
- 1 vector por cada audio en el conjunto de entrenamiento inicial.
- La salida esperada es la emoción etiquetada para el audio correspondiente

## Capas ELM

- Capa entrada: 4 unidades por cada emoción
- Capa oculta: 120 unidades
- Salida: Vector  $K$ -dimensional, con las  $K$  probabilidades



# Experimentos

- Datos: Contenido audiovisual de 10 actores.
- Cada Audio fue etiquetado con una emoción.
  - Exaltación (*Excitement*)
  - Frustración
  - Felicidad
  - Neutro
  - Sorpresa
- No aprovechan información del hablante.
- $\theta = 0.2$

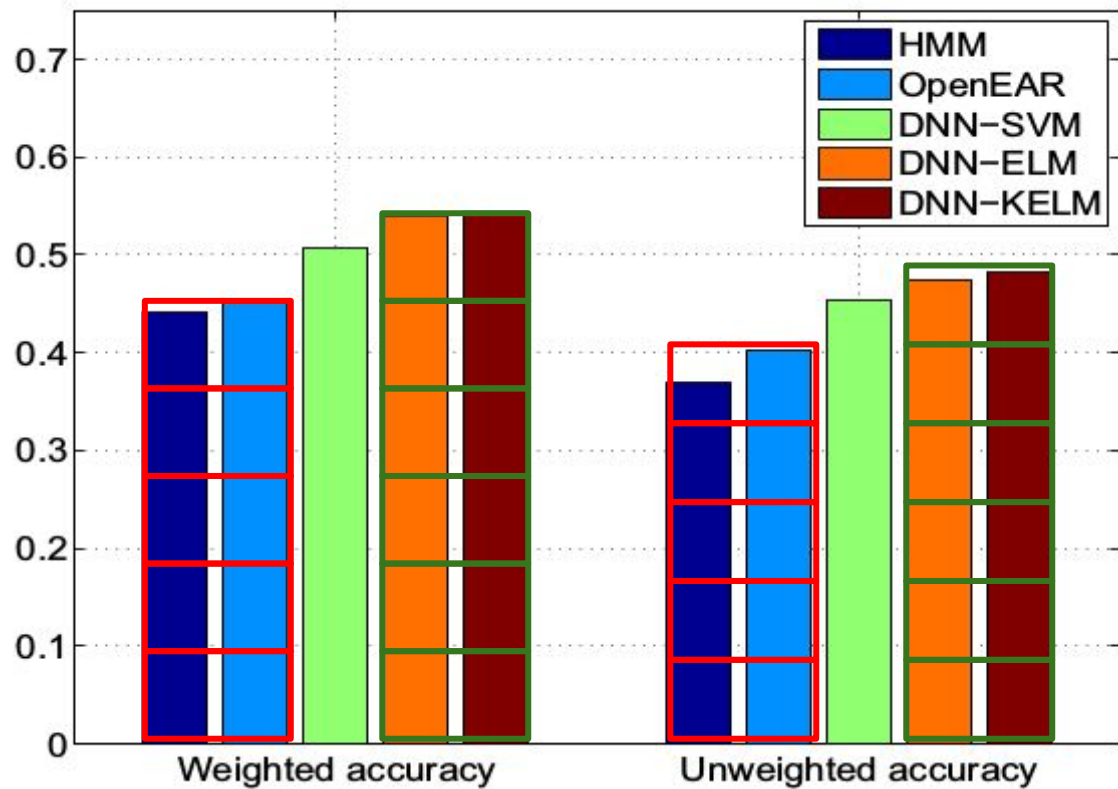




# Experimentos

- Comparan contra
  - Enfoque con HMMs y atributos sobre tono y amplitud.
  - OpenEAR que usa SVM con atributos estadísticos
  - Su propio enfoque pero reemplazando la ELM por
    - SVM
    - Kernel ELM
- Metricas
  - *Weighted Accuracy*
    - Precisión de clasificación de la emoción sobre el total del conjunto de prueba
  - *Unweighted Accuracy*
    - Promedio de las precisiones por cada emoción.





Comparación de resultados para los distintos enfoques