Tesi di laurea in Sistemi ad Agenti

EmoFEATURES:

RICONOSCIMENTO DELLE EMOZIONI DALLA VOCE

Relatore: Prof.ssa Berardina De Carolis

Laureando: Michele Metta

Anno Accademico 2018/2019

dib

Obiettivi

- Valutazione di algoritmi di apprendimento automatico e deep learning con corpora di voci emotive multilingua (italiano,inglese, spagnolo e tedesco) per il riconoscimento delle emozioni dalla voce mediante valence e arousal.
- Creazione di un sistema (EmoFEATURES) per il riconoscimento delle emozioni in real time.

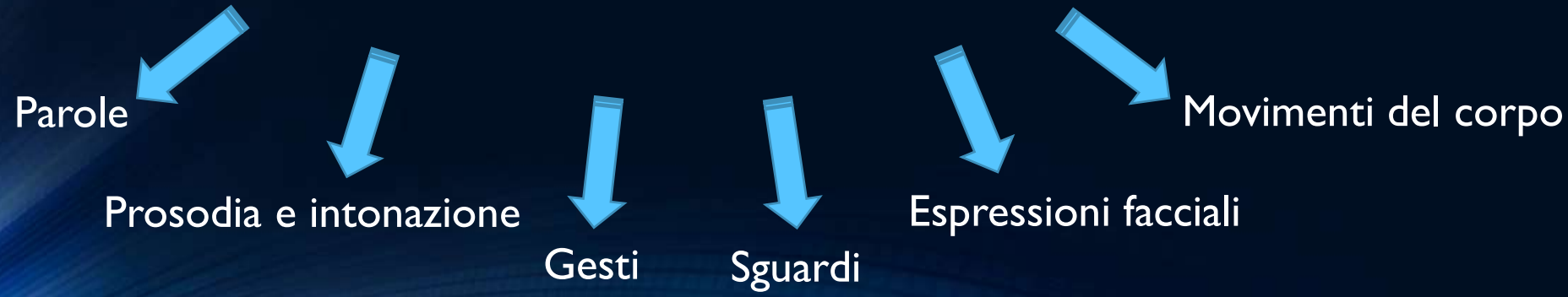
Riconoscimento delle Emozioni

EMOZIONI:

- *Processo multicomponenziale, articolato in più componenti, le emozioni hanno un decorso temporale e sono attivate da stimoli interni o esterni.*
- Caratteristica distintiva dell'uomo
- Rapporti interpersonali basati sulle emozioni

FUNZIONE COMUNICATIVA:
L'emozione comunica all'esterno lo stato cognitivo e le intenzioni di una persona. Questa comunicazione può avvenire sia in modo consapevole che inconsapevole.

Il nostro corpo dispone di vari sistemi di comunicazione



Categorizzare l'emozione (1)

Come formalizzare l'emozione?

Negli anni si è cercato di categorizzare l'emozione umana, al fine di formalizzarla.

Creare un modello, *unico*, per categorizzare tutte le sfaccettature delle emozioni è **impossibile.**

I modelli maggiormente riconosciuti sono due:

1. Modello **discreto**
2. Modello **dimensionale**

Categorizzare l'emozione (2)

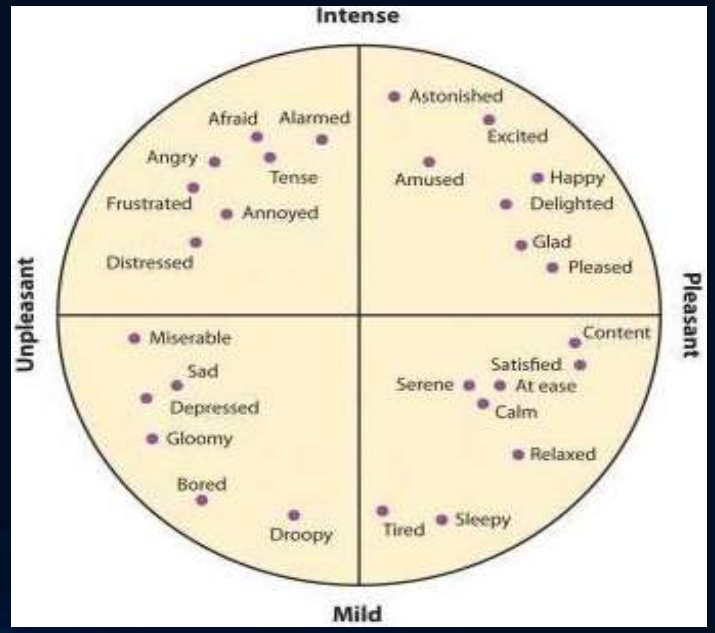
Modello **discreto**

- Teoria formulata da *Paul Ekman* negli anni '70

Modello **dimensionale**

(Modello circonflesso delle emozioni)

- Definito da *James Russell* nel 1980
- *Definisce gli stati emotivi con una coppia: valence* (positiva, negativa, neutrale) e *arousal (alta, bassa, media).*



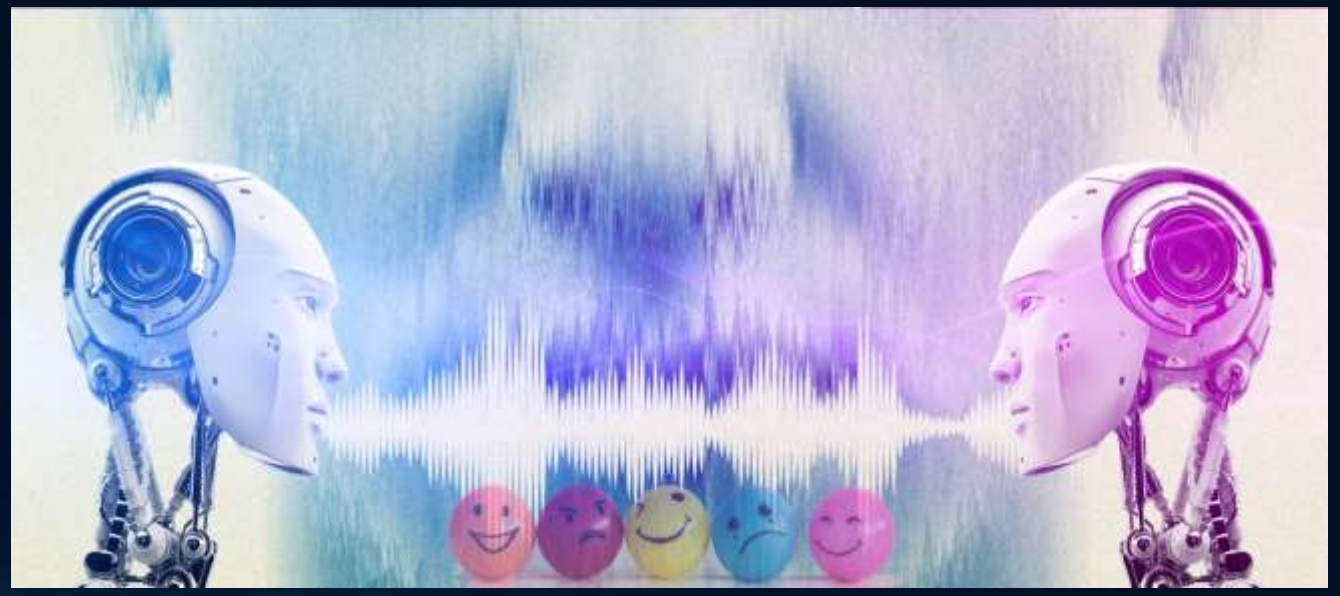
Emozioni

Emozione	Valence	Arousal
Disgusto	Negativa	Bassa
Gioia	Positiva	Alta
Paura	Negativa	Alta
Neutrale	Neutrale	Media
Rabbia	Negativa	Alta
Sorpresa	Positiva	Alta
Tristezza	Negativa	Media

Speech Emotion Recognition

Si definisce *Speech Emotion Recognition* (SER) la capacità di riconoscere lo stato emotivo di una persona dalla voce.

- I servizi di riconoscimento delle emozioni sono molto limitati.
- A differenza dello speech e text recognition non esistono modelli efficaci.
- Emozioni vocali: più soggettive, non è semplice misurarle e categorizzarle oggettivamente.
- Classificazione su 3 classi: miglioramento risultati, ambienti più utile la valence rispetto all'arousal, o viceversa.



Datasets (1)

- **Ravdess** (Ryerson Audio-Visual Database of Emotional Speech and Song)
    - 1440 file audio (48 kHz in formato wav), 24 attori (12 di sesso maschile, 12 di sesso femminile), lingua inglese.
- **Tess** (Toronto emotional speech set)
    - 2800 file audio (48 kHz), prima attrice 26 anni, seconda attrice 64 anni, lingua inglese.
- **Savee** (Surrey Audio-Visual Expressed Emotion)
    - 480 audio (44 kHz), 4 oratori maschili, 15 frasi per ogni emozione (neutralità 30 frasi), inglese.
- **Emofilm**
    - Dataset multilingua (italiano, inglese, spagnolo).
    - 1115 audio (48 kHz in wav, monocanale) registrati da 43 film.

Datasets (2)

- **Emovo**
    - Lingua italiana, 588 audio (48 kHz, formato wav), 6 attori (3 uomini, 3 donne).
- **Demos** (Database of Elicited Mood in Speech)
    - Lingua italiana, 8568 file audio totali (48 kHz, formato wav).
    - Utilizzato per il bilanciamento.
- **Emodb** (Berlin Database of Emotional Speech)
    - 494 audio, registrati in una camera anecoica, 10 attori professionisti (5 uomini, 5 donne), tedesco.

Pre-elaborazione audio

- Frequenza di campionamento omogenea a 16 kHz.
- Rimozione silenzio.
- Audio al di sotto di 1 secondo eliminati (poche informazioni).
- Data augmentation (Rumore Gaussiano - maggiore generalizzazione dei modelli e riduzione overfitting)

Dataset bilanciato (i modelli tendono a concentrarsi sulle classi maggioritarie, le osservazioni che appartengono alle classi minoritarie vengono classificate erroneamente più facilmente).

Emozione	Numero audio
Disgusto	2060
Gioia	2060
Neutrale	2060
Paura	2060
Rabbia	2060
Sorpresa	2060
Tristezza	2060

Totale: 14420 file audio

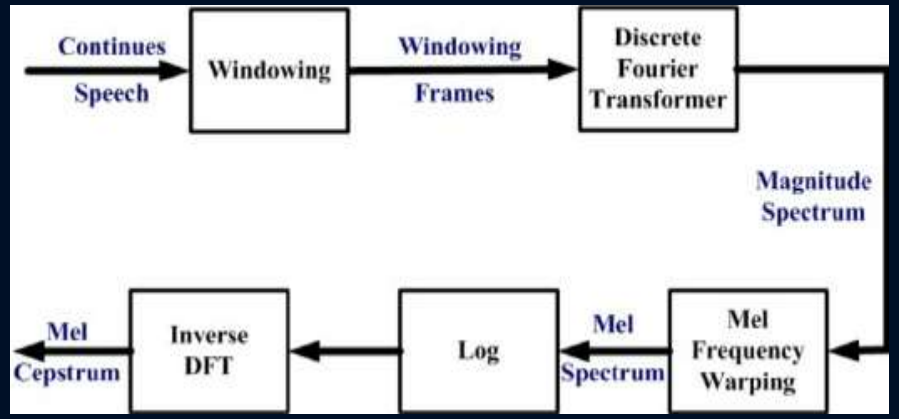
Features utilizzate (1)

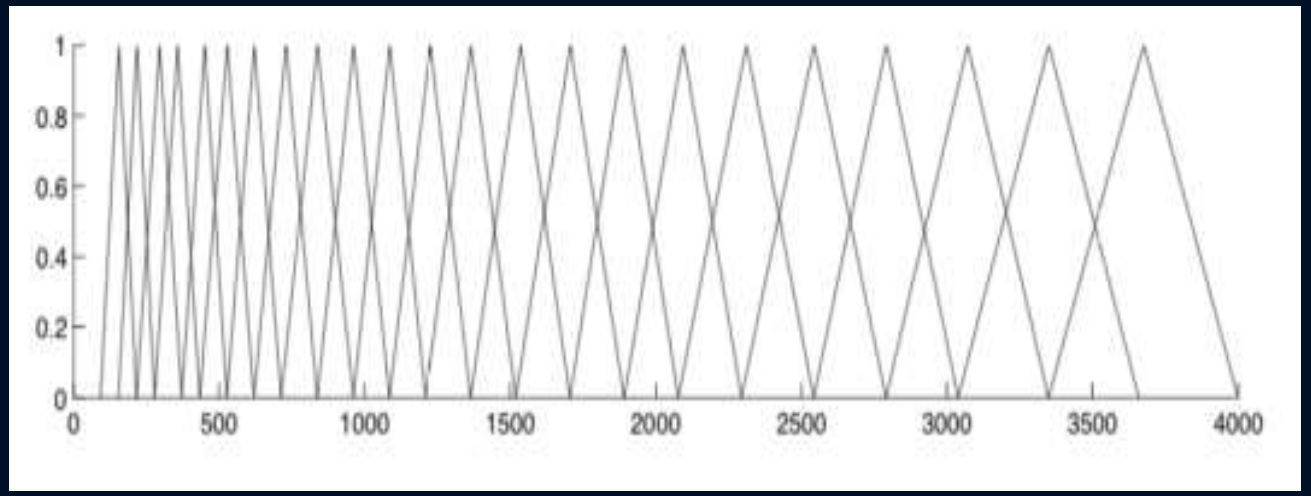
Mfcc (Mel-Frequency-Cepstral-Coefficient)

- Simulano l'orecchio umano: numero di recettori distribuiti in maniera logaritmica esternamente (alte frequenze), lineare nella parte interna (basse frequenze).

- Utilizzano il Filtro Mel per trasformare i dati sonori in modo lineare alle basse frequenze (sotto i 1000 Hz), logaritmico alle alte (sopra i 1000 Hz)

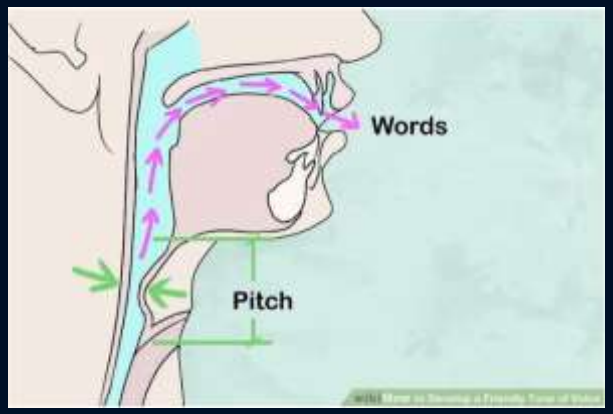
- 20 coefficienti utilizzati.
  - A partire dai 20 coefficienti sono state calcolate delta e delta-delta (andamento temporale).



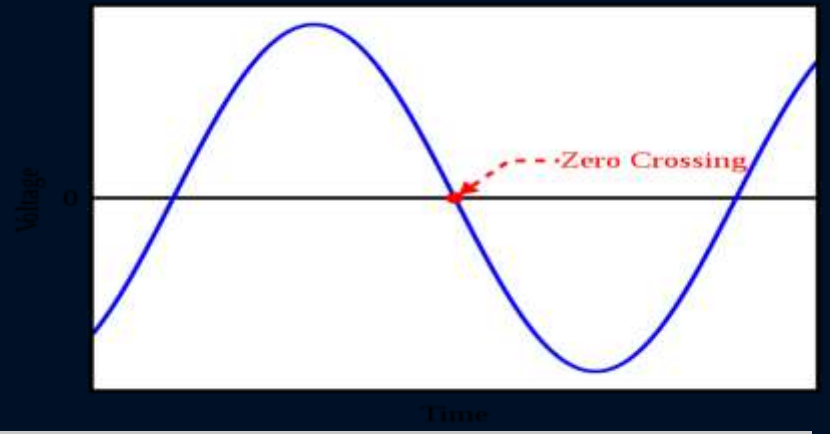


Features utilizzate (2)

Pitch
- Corrisponde all'intonazione della voce (frequenza di vibrazione delle corde vocali).
- Livelli alti: gioia, rabbia, paura (mirano a catturare l'attenzione di uno o più ascoltatori).
- Livelli bassi: tristezza, calma (emozioni più sobrie).

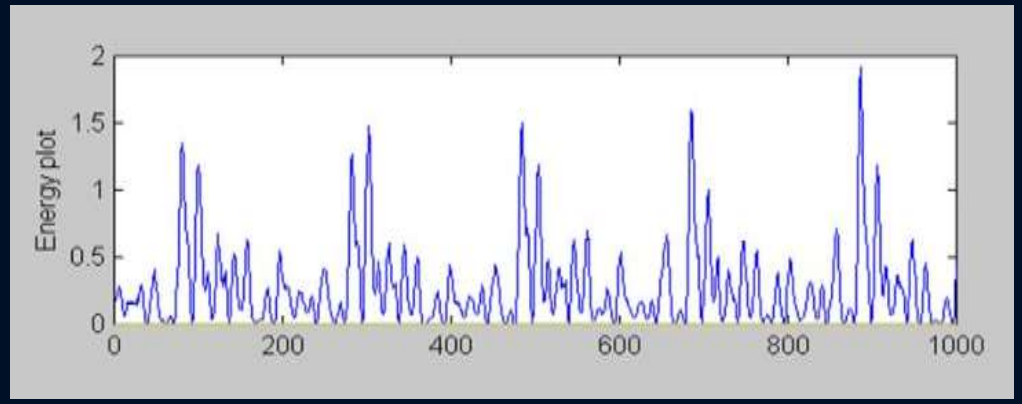


ZCR (Zero Crossing Rate)
- Velocità di attraversamento dello zero (misura la velocità delle variazioni di segno all'interno di un segnale audio, da positivo a zero a negativo o da negativo a zero a positivo).



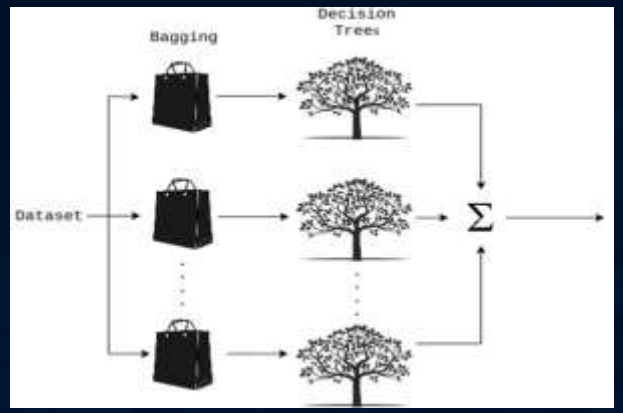
Energy

- Permette di calcolare l'intensità e quindi la potenza del suono emesso. (intensità crescente per emozioni come rabbia, paura, gioia; intensità decrescente per emozioni come la tristezza.

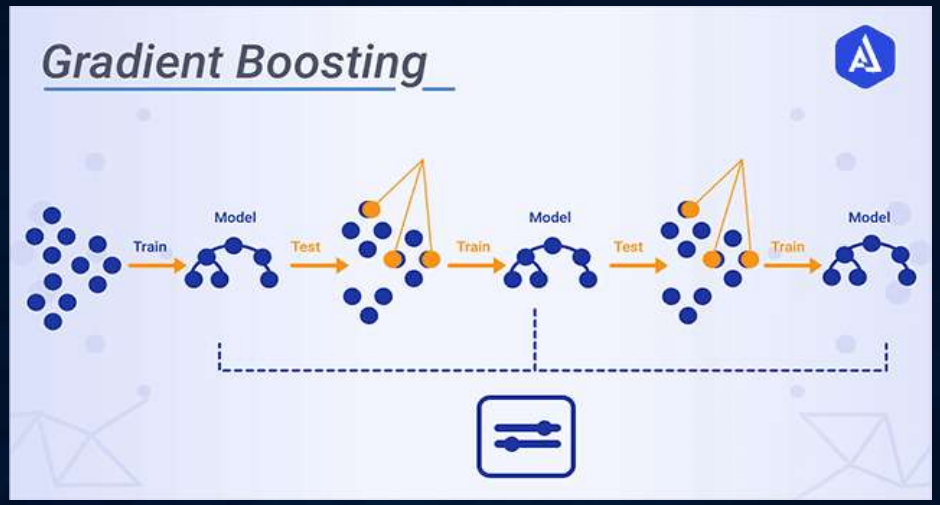


Classificatori

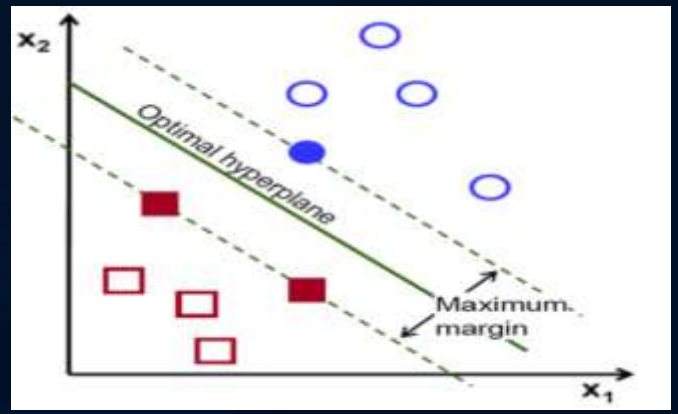
Random Forest



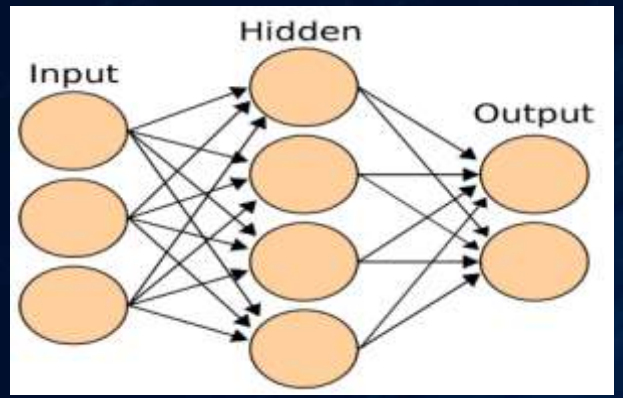
Gradient Boosting



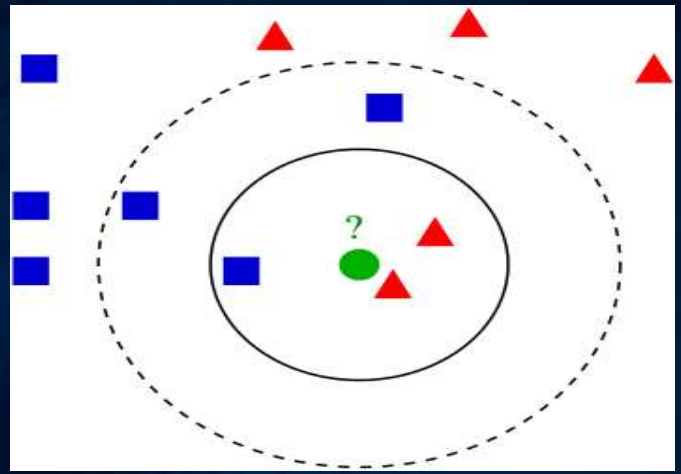
Support Vector Machine



Multi–Layer Perceptron



K-Nearest-Neighbors



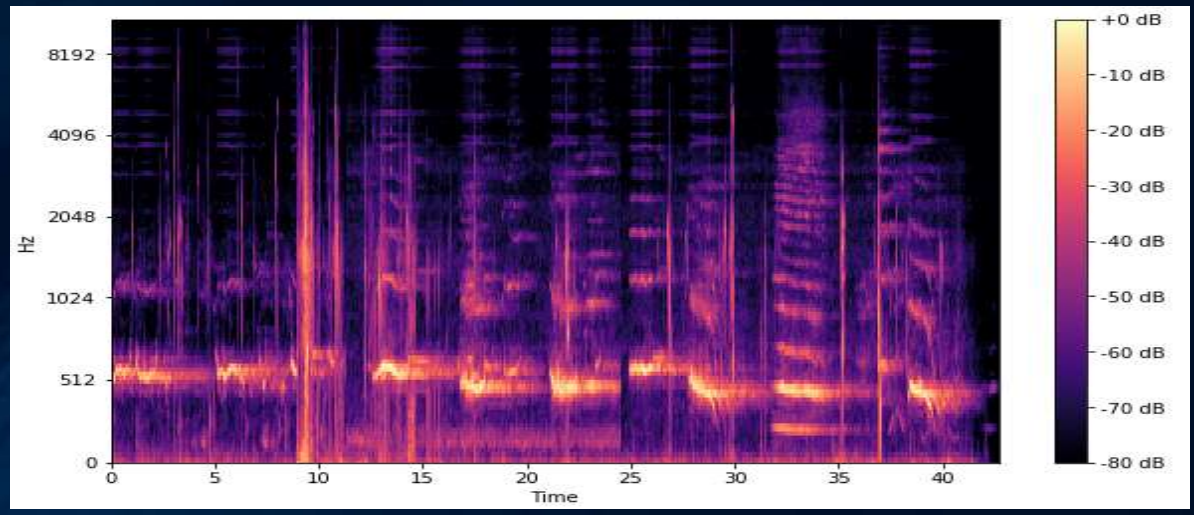
Valori features standardizzati per ottenere media 0 e varianza 1.

Convolutional Neural Network (CNN)

- Input:
  - Estrazione del Mel-Frequency Spectrogram: descrive come il contenuto frequenziale varia nel tempo.

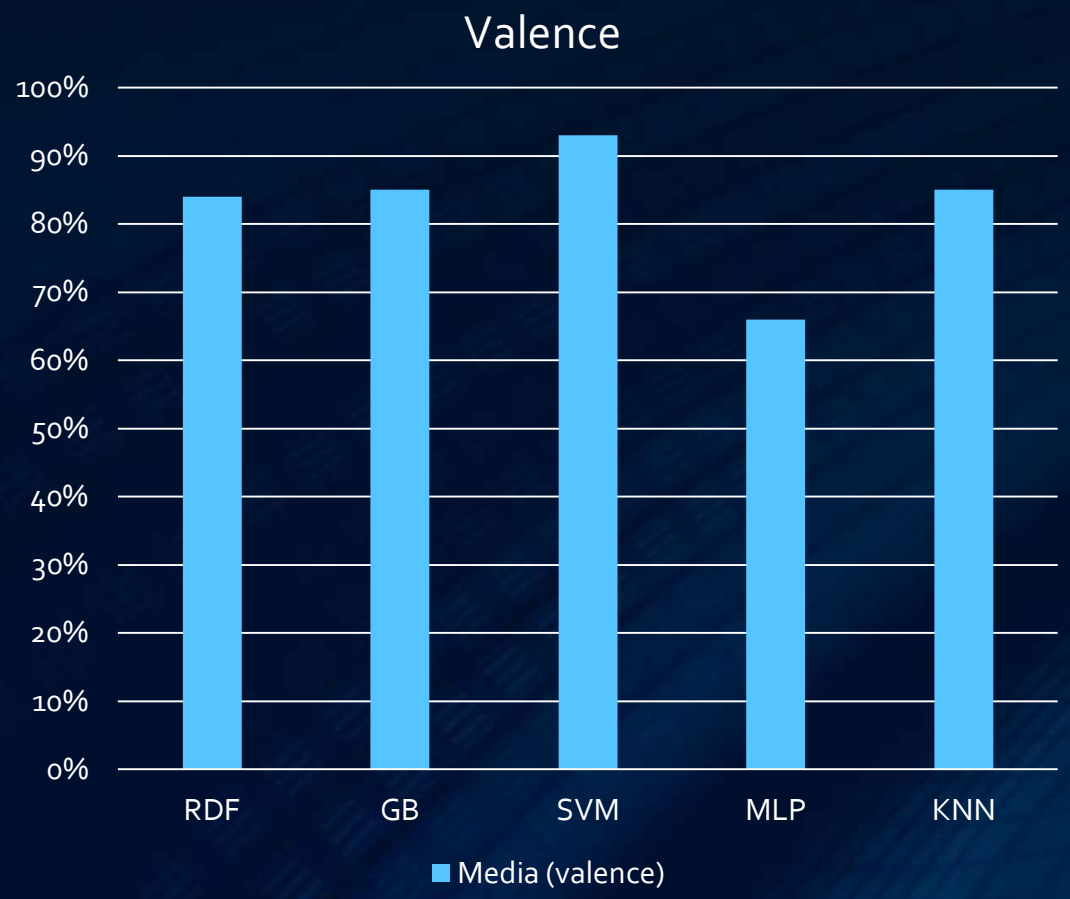
    - Sull'asse y la frequenza è riportata in scala Mel. Questo pone maggiore attenzione sull'estremità inferiore dello spettro delle frequenze rispetto a quello superiore, imitando così le capacità percettive dell'udito degli umani. Sull'asse x invece continua ad essere riportato il tempo in scala lineare.

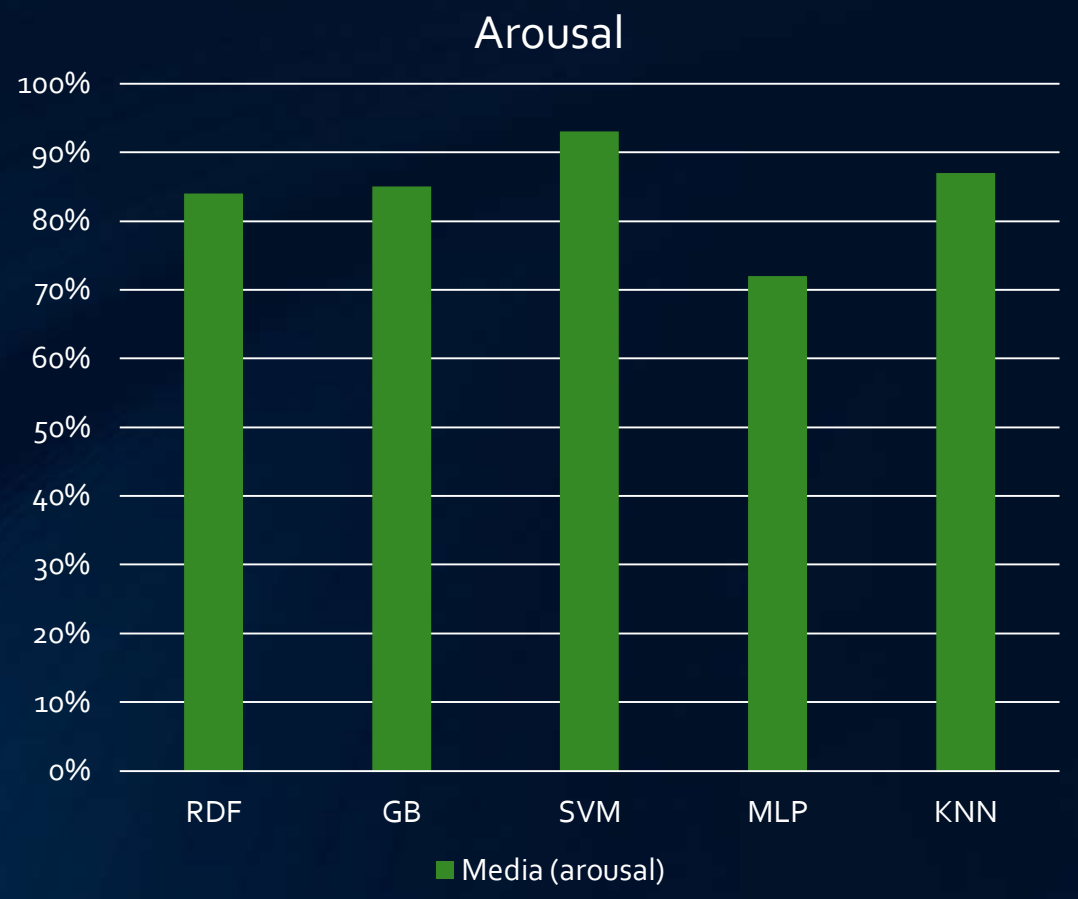
  - Dimensione spettrogramma: (128,128) - 3 secondi.
    - Audio corti: zero-padding.
    - Audio lunghi: troncamento.



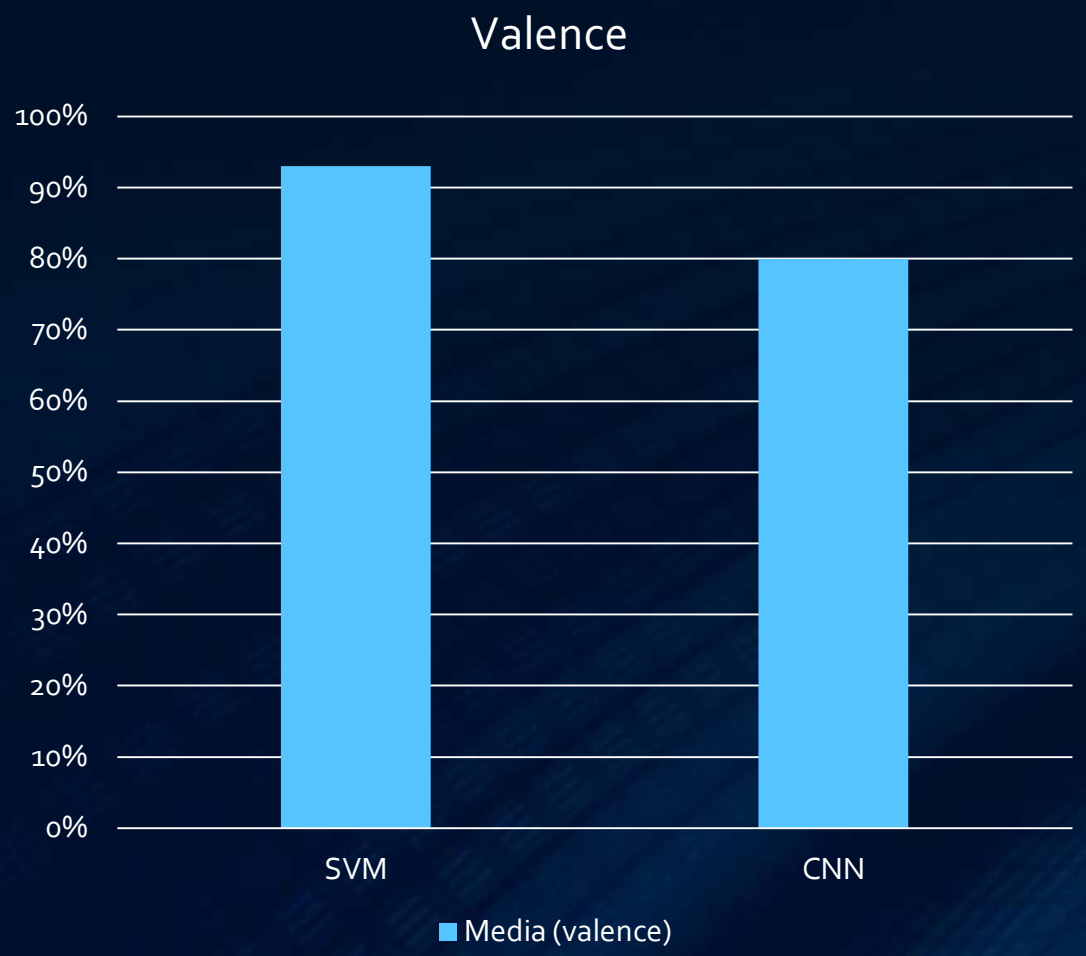
Sperimentazione e risultati (dataset bilanciato)

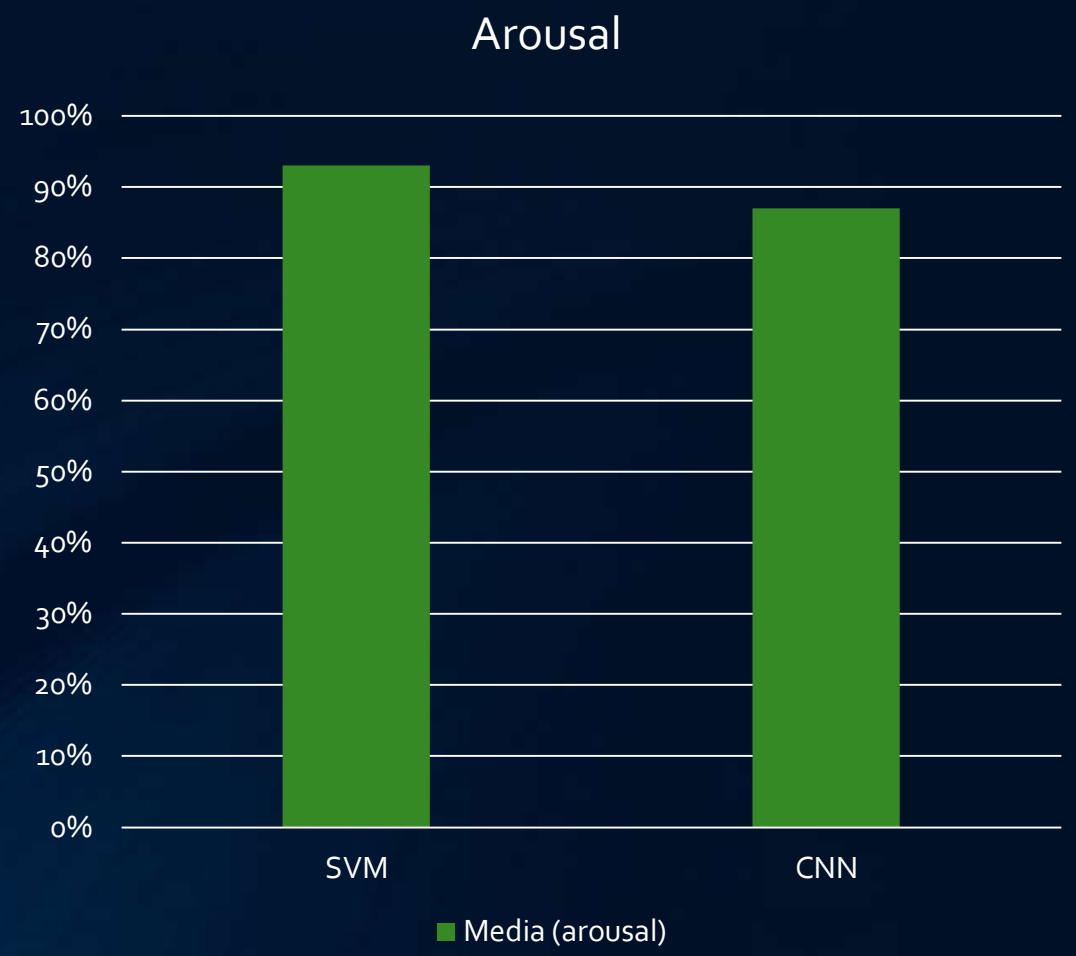
- Stratified k-fold cross-validation (stessa percentuale di campioni per ogni partizione) - **k = 5**.
- Valori di accuratezza su partizioni equilibrate che garantiscono una buona rappresentazione di tutte le classi.





SVM vs CNN (dataset bilanciato)

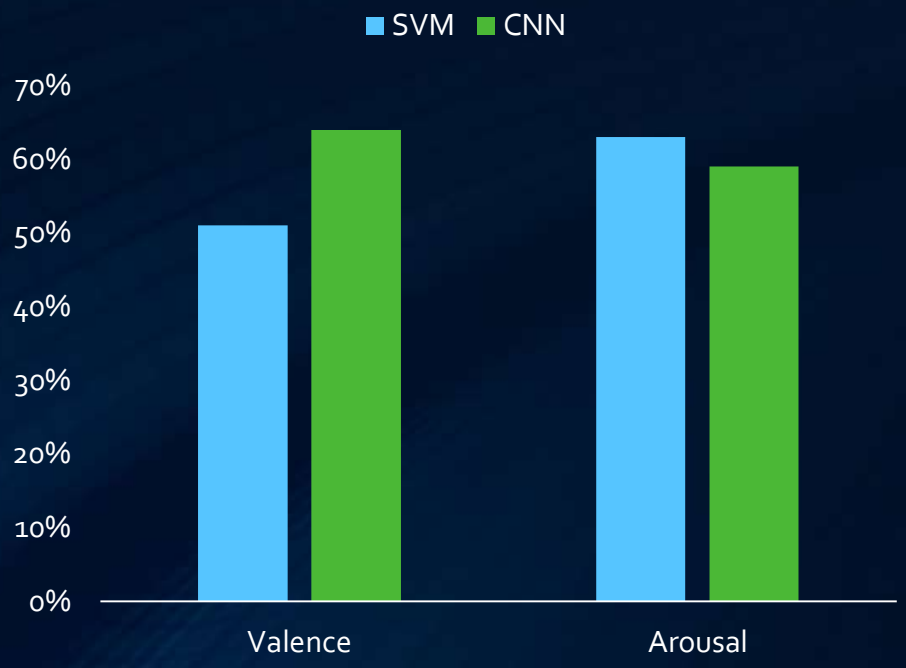




Risultati in the wild

- 209 file audio estratti da film.
- Etichettati da 7 persone.

Emozione	Valence	Arousal	Num. Audio
Disgusto	Negativa	Bassa	30
Gioia	Positiva	Alta	30
Paura	Negativa	Alta	30
Neutrale	Neutrale	Media	29
Rabbia	Negativa	Alta	38
Sorpresa	Positiva	Alta	22
Tristezza	Negativa	Media	30

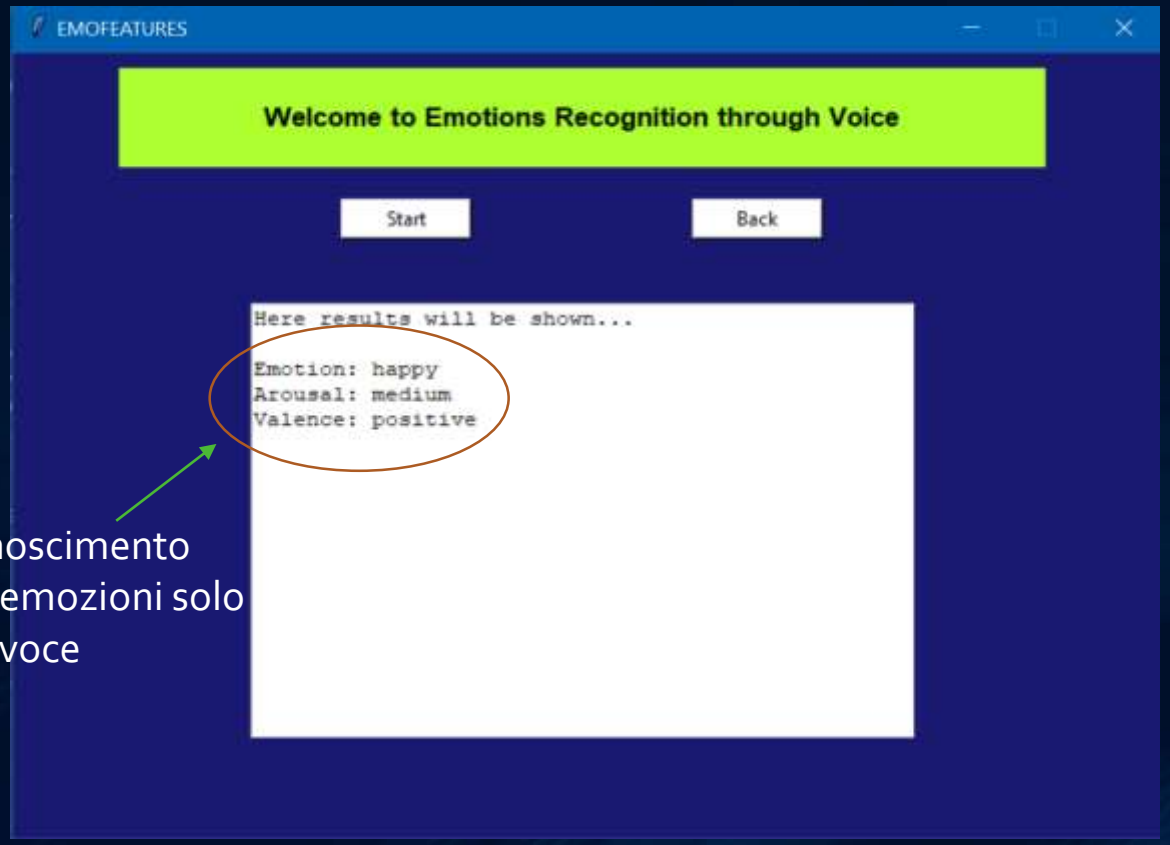


Valence:
SVM: 51%
CNN: 64% (+ 13%)

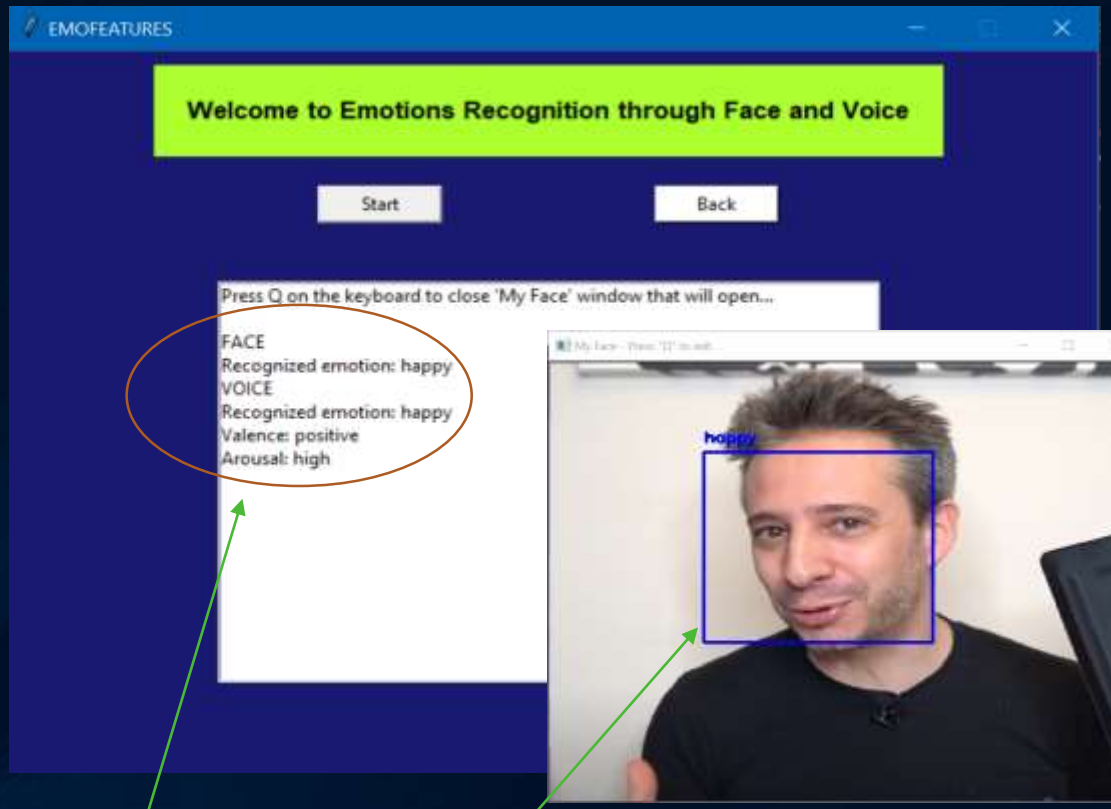
Arousal:
SVM:63% (+ 4%)
CNN:59%

Modelli finali utilizzati:
Valence: CNN
Arousal: SVM

EmoFEATURES



Riconoscimento delle emozioni solo dalla voce



Riconoscimento delle emozioni dalla voce e dal volto contemporaneamente

Conclusioni e sviluppi futuri

Conclusioni:

- SVM miglior modello di classificazione tra quelli testati sul dataset bilanciato per la valence e l'arousal.
- Sul dataset in the wild: CNN in grado di generalizzare meglio la valence (64%, +13%), SVM ancora migliore per l'arousal (63%, +4%).
- Valori di accuratezza bassi dovuti all'incidenza di diversi fattori: ambiente non controllato, strumentazioni non professionali, emozioni soggettive, dipendenza dal genere (maschile, femminile).

Sviluppi futuri:

- Aumentare dimensione del dataset (Le CNN lavorano meglio se addestrate su grandi quantità di dati).
- Suddividere il dataset per genere, in modo da valutare eventuali influenze dovute alla differenza di genere degli oratori.
- Implementazione di una funzione di riaddestramento per la voce in EmoFEATURES (riaddestramento modelli per imparare da eventuali esempi erroneamente predetti).
- Individuare una tecnica per poter fondere i risultati della voce e del volto in modo da cercare di ottenere risultati più accurati.

Grazie per l’attenzione