UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

DOCUMENTAZIONE PROGETTO DL & NLP

Fairness Speech Recognition

TEAM

Simona Lo Conte Matricola: 0522501786

Marta Napolillo Matricola: 0522501787

Anno Accademico 2024-2025

Abstract

Il presente lavoro affronta il problema della fairness nei sistemi di riconoscimento automatico del parlato (ASR), con l'obiettivo di individuare eventuali disparità nelle performance dei modelli, in relazione a genere e dialetto degli speaker. Dopo una panoramica generale sui principi e sul funzionamento dei sistemi ASR, viene proposta un'analisi dello stato dell'arte, con particolare attenzione ai più recenti studi sul bias nei modelli di riconoscimento vocale, che evidenziano come alcune categorie di parlanti possano essere sistematicamente svantaggiate in fase di trascrizione automatica. La metodologia prevede l'utilizzo del dataset `spotify_podcast_ASR`, ricco di metadati sociolinguistici, su cui vengono applicate due strategie principali: la progettazione di modelli RNN (LSTM e GRU) e l'inferenza tramite modelli preaddestrati della libreria HuggingFace (Whisper-medium multilingual e Wav2Vec2-base). Tuttavia, a causa delle elevate risorse computazionali richieste per l'addestramento delle RNN, la fase di esecuzione completa di questi modelli è stata rimandata a sviluppi futuri. L'analisi è stata quindi concentrata sui modelli preaddestrati, valutandone le prestazioni in base al genere, al dialetto, e alla combinazione dei due fattori, con l'obiettivo di individuare la presenza di bias sistematici. I risultati mostrano che il modello Whisper-medium multilingual ottiene prestazioni superiori rispetto a Wav2Vec2-base, con valori di WER e CER più bassi. Questo miglioramento è dovuto alla natura multilingue del modello Whisper, che gli consente di gestire con maggiore flessibilità la varietà linguistica presente nel dataset. Per quanto riguarda la variazione dialettale, il dialetto SAE (Standard American English) registra le performance migliori in termini di WER, il che può essere attribuito alla sua maggiore rappresentazione all'interno del dataset, che favorisce l'adattamento del modello a tale varietà. Infine, l'analisi di genere non evidenzia significative disparità tra uomini e donne; tuttavia, si osserva una leggera tendenza a migliori prestazioni con le voci femminili.

Indice

CAPITOLO 1

Introduzione

1.1 Automatic Speech Recognition (ASR)

Il riconoscimento automatico del parlato (Automatic Speech Recognition, ASR) è l'ambito della ricerca che si occupa di trasformare il linguaggio parlato in testo scritto in modo automatico. Questo processo richiede la capacità di interpretare correttamente segnali audio, distinguere le parole pronunciate e ricostruire frasi coerenti e grammaticalmente corrette.

L'ASR è un campo in continua evoluzione e di fondamentale importanza per una vasta gamma di applicazioni pratiche, tra cui assistenti vocali, sistemi di trascrizione automatica, strumenti di sottotitolazione e applicazioni rivolte all'accessibilità. Il riconoscimento del parlato implica la comprensione non solo dei suoni prodotti, ma anche delle loro variazioni dovute a fattori come il ritmo, l'intonazione, gli accenti e le inflessioni personali.

Poiché la comunicazione verbale è estremamente variabile da persona a persona, uno dei principali obiettivi dell'ASR è garantire una trascrizione accurata e affidabile, indipendentemente dalle caratteristiche del parlante o dal contesto comunicativo.

Per questa ragione, la qualità dei dati audio, la varietà linguistica rappresentata e la robustezza del sistema giocano un ruolo essenziale nel determinare l'efficacia del riconoscimento.

1.2 Fairness

Il concetto di *fairness* nell'ambito dell'intelligenza artificiale si riferisce alla capacità di un sistema di operare in modo equo nei confronti di tutti gli individui, senza introdurre bias o discriminazioni legate a caratteristiche personali o demografiche.

1.2.1 Fairness in ASR

All'interno dei sistemi di ASR, la fairness rappresenta un criterio cruciale da considerare, poiché le prestazioni del sistema possono variare sensibilmente in base al genere, all'accento, all'appartenenza etnica o ad altri fattori sociolinguistici. In questo lavoro ci concentriamo quindi sull'analizzare le implicazioni etiche e sociali in merito al task di riconoscimento automatico del parlato.

1.3 Struttura del documento e obiettivi del progetto

Questo progetto si propone di analizzare potenziali bias sistematici nei modelli di riconoscimento automatico del parlato (ASR), utilizzando il dataset *spotify_podcast_ASR*, che include ricchi metadati sociolinguistici. L'approccio adottato prevede due strategie principali: da un lato, la progettazione di modelli RNN (basati su LSTM e GRU), e dall'altro, l'impiego di modelli preaddestrati disponibili nella libreria HuggingFace, quali Whisper-medium multilingual e Wav2Vec2-base. L'analisi condotta si è focalizzata sui modelli preaddestrati, valutandone le prestazioni in relazione al genere, al dialetto e alla loro combinazione, al fine di evidenziare eventuali disparità o bias presenti nel riconoscimento del parlato.

Il documento è strutturato in più sezioni. Nella Sezione 2 c'è un'analisi esplorativa dei lavori correlati alla problematica in esame. Si passa poi alla Sezione 3, in cui viene

descritta in maniera dettagliata la metodologia condotta per la creazione del sistema di ASR tramite RNN e tramite modelli preaddestrati. La Sezione 4 contiene un'analisi dei risultati ottenuti, per poi passare alla Sezione 5 che comprende le conclusioni ed eventuali sviluppi futuri, di cui poter tener conto in merito alla tematica affrontata.

CAPITOLO 2

Stato dell'arte

2.1 ASR e Fairness

Negli ultimi anni, il riconoscimento automatico del parlato (ASR) ha compiuto significativi progressi grazie all'impiego di tecniche di DL e alla disponibilità di grandi quantità di dati audio. Tuttavia, parallelamente ai miglioramenti in termini di accuratezza, è emersa una crescente attenzione verso le implicazioni etiche e sociali di tali sistemi, in particolare rispetto al loro comportamento nei confronti di gruppi eterogenei di utenti.

Diversi studi hanno mostrato come i sistemi ASR possano favorire in modo implicito i gruppi linguistici più rappresentati nei dati, penalizzando invece i parlanti appartenenti a minoranze o con caratteristiche vocali meno comuni.

In questa sezione vengono analizzati i principali studi relativi all'equità nei sistemi ASR, evidenziando le sfide, le metriche utilizzate e gli approcci proposti per valutare e mitigare i bias presenti nei modelli.

2.2 Analisi dello stato dell'arte

Un contributo significativo allo studio della fairness nei sistemi di riconoscimento vocale è fornito dal lavoro di Veliche et al. [1], che introduce il dataset Fair-Speech, progettato specificamente per valutare l'equità delle performance nei modelli ASR. Il corpus comprende circa 26.500 enunciati vocali, registrati da 593 partecipanti residenti negli Stati Uniti, che hanno fornito anche informazioni demografiche auto-riferite, tra cui età, genere, etnia, provenienza geografica e lingua madre. Oltre al rilascio dei dati, gli autori forniscono una valutazione comparativa basata su diversi modelli ASR, tra cui modelli supervisionati, semi-supervisionati e il modello Whisper, evidenziando consistenti differenze nei tassi di errore (WER) tra i gruppi demografici.

Il lavoro condotto da Xia et al. [2] si focalizza sui dataset per hate speech recognition, notando che spesso AAE viene etichettato erroneamente come offensivo o di odio, portando a un'elevata percentuale di falsi positivi durante la classificazione di discorsi d'odio. Pertanto, essi presentano un approccio basato su adversarial training, ponendo AAE come protected attribute e prevedendo l'attributo tossicità come target, al fine di mitigare il rischio di pregiudizi razziali nei classificatori di discorsi d'odio, anche in presenza di annotazioni errate.

Liu et al. [3] si concentrano sul corpus Casual Conversation, contenente oltre 800 ore di parlato spontaneo annotato con metadati sensibili come genere, età e tonalità della pelle. In una prima fase il loro lavoro si è concentrato sulla trascrizione manuale di tale dataset, al fine di poterlo utilizzare per valutare eventuali bias nei modelli. A seguito di tale fase, viene condotta un'analisi della fairness per il task di riconoscimento vocale. Lo studio evidenzia significative differenze di accuratezza nei modelli ASR analizzati, con tassi di errore generalmente più bassi per i parlanti di genere femminile. Sebbene il colore della pelle non impatti direttamente il parlato, gli autori mostrano come esso possa correlare con altri fattori che influenzano le prestazioni del riconoscimento vocale, contribuendo così all'analisi della fairness nei sistemi ASR.

L'obiettivo di Harris et al. [4] è quello di valutare le prestazioni dei sistemi ASR tra i diversi generi e tra i dialetti della lingua inglese. Per fare ciò, partono dalla costruzione di un set di dati etichettati di 13 ore di podcast, trascritti da parlanti dei dialetti rappresentati. A seguito della costruzione del corpus, passano alla valutazione delle prestazioni confrontando diversi modelli di ASR (Wav2Vec2, Whisper e huBERT in diverse versioni) in termini di Word Error Rate ottenuta. Da tale analisi comparativa emerge che il modello migliore è Whisper-medium. Questo studio rappresenta un contributo rilevante nel campo della fairness applicata al riconoscimento vocale, fornendo una base solida per future ricerche orientate alla riduzione delle disparità tra gruppi demografici.

CAPITOLO 3

Metodologia

3.1 Panoramica generale

In questo capitolo viene descritta nel dettaglio la metodologia adottata per lo sviluppo del progetto, articolata in diverse fasi. Si parte da un'analisi preliminare del dataset, volta a comprenderne la struttura, le features principali e la distribuzione dei dati, per poi passare all'implementazione dei modelli Automatic Speech Recognition (ASR).

Nella prima fase sperimentale, sono state progettate due architetture basate su Recurrent Neural Network (RNN): Long Short-Term Memory (LSTM) e Gated Recurrent Unit (GRU). L'obiettivo era valutare le capacità di queste reti nell'affrontare il compito di ASR, sfruttando la loro capacità di modellazione di sequenze temporali.

Tuttavia, considerate le migliori performance dei modelli basati su architetture Transformer nella letteratura recente, è stata successivamente intrapresa una seconda fase sperimentale. Nello specifico, sono stati utilizzati due modelli preaddestrati messi a disposizione dalla libreria HuggingFace: Whisper-medium multilingual e Wav2Vec2-base, entrambi noti per le loro elevate prestazioni nel campo dell'ASR.

Le sezioni successive approfondiscono ciascuna delle fasi implementative, illustrando le scelte tecniche effettuate, le configurazioni dei modelli e le strategie adottate per la valutazione delle performances.

3.2 Research Questions

Al fine di guidare e analizzare al meglio i risultati prodotti in tale progetto, vengono definite in una fase preliminare le cosiddette domande di ricerca. Le research questions definiscono gli obiettivi specifici dell'analisi, orientando la progettazione della pipeline e la scelta delle metodologie adottate. In particolare, nel nostro caso specifico, l'attenzione è rivolta alla valutazione della fairness nei sistemi di riconoscimento vocale automatico, con un focus su possibili disparità nelle performance tra diversi gruppi demografici. Per chiarire al meglio i risultati ottenuti, sono state formulate le seguenti tre domande di ricerca.

Le domande di ricerca si concentrano sulle performance ottenute con i modelli Whisper-medium multilingual e Wav2Vec2-base, mentre l'addestramento delle RNN è stato rimandato a sviluppi futuri, poiché troppo oneroso dal punto di vista computazionale.

🔍 **RQ$_1$.** *Quale architettura di rete neurale dei modelli preaddestrati — Whisper-medium multilingual e Wav2Vec2-base — garantisce le migliori prestazioni in termini di Word Error Rate (WER) e Character Error Rate (CER) sull'intero dataset?*

La prima domanda di ricerca si propone di identificare quale tra i modelli sviluppati fornisce le migliori prestazioni. A tal fine, vengono analizzate le metriche WER e CER, descritte in dettaglio nella Sezione 3.6. La valutazione è stata condotta esclusivamente in fase di inferenza.

🔍 **RQ$_2$.** *In quali gruppi sociolinguistici (definiti da dialetto e genere separatamente) si osservano i più alti tassi di errore, suggerendo la presenza di bias sistematici nei modelli ASR?*

La seconda domanda di ricerca si focalizza sull'identificazione dei bias presenti nei modelli di ASR. In particolare, l'analisi sfrutta le informazioni relative al genere e al dialetto dei parlanti per individuare i contesti sociolinguistici in cui si registrano i tassi di errore più elevati, misurati tramite le metriche WER e CER. Tale valutazione è stata condotta in fase di inferenza per i modelli preaddestrati.

RQ$_3$. *Quale dialetto mostra il più basso gender bias in termini di performance, in modo ricorrente su tutti i modelli ASR valutati?*

La terza domanda di ricerca mira a identificare, per ciascun modello ASR considerato, il dialetto che presenta il minor gender bias. A tal fine, per ogni varietà dialettale si analizzano le differenze nei tassi di errore tra i generi, utilizzando le metriche WER e CER. L'analisi è stata condotta in fase di inferenza per i modelli preaddestrati.

3.3 Dataset

Per lo sviluppo di tale progetto è stato fatto riferimento al paper [4], descritto in dettaglio nel Capitolo 2. Il dataset scelto in tale analisi è lo stesso utilizzato dall'analisi di Harris et al. In particolare, il dataset è disponibile sulla piattaforma HuggingFace (link al dataset) e accessibile tramite il seguente snippet di codice.

```
from datasets import load_dataset
ds = load_dataset("SALT-NLP/spotify_podcast_ASR")
```

Il dataset `spotify_podcast_ASR` è stato appositamente creato per supportare lo sviluppo e la valutazione di sistemi di Automatic Speech Recognition (ASR) applicati a contenuti audio tipici dei podcast. Esso comprende un totale di 1690 elementi, rappresentati dai seguenti attributi rilevanti:

- **audio**, contenente i dati audio dei podcast;
- **Unnamed: 0**, rappresentante l'identificatore univoco associato a ciascuna istanza;

- **transcription**, ossia la trascrizione testuale dell'audio corrispondente, utilizzata come riferimento per la valutazione dei modelli ASR;
- **aave**, variabile binaria indicante se l'audio contiene persone con accento di tipo African American Vernacular English (AAVE);
- **chicano_english**, variabile binaria che segnala la presenza di caratteristiche linguistiche del Chicano English;
- **spanglish**, variabile binaria che indica la presenza di dialetto Spanglish;
- **sae**, variabile binaria segnalante che il parlato è conforme allo Standard American English (SAE);
- **other_dialect_accent**, indica la presenza di altri accenti o dialetti non esplicitamente categorizzati negli altri attributi;
- **codeswitching**, indicante se nel segmento audio è presente o meno un passaggio tra più lingue (tipicamente tra inglese e spagnolo);
- **women**, variabile binaria che indica se il parlante è una donna;
- **men**, variabile binaria che indica se il parlante è un uomo.

Per ogni dialetto e genere considerato nel dataset, viene riportato anche il numero di parlanti all'interno dell'audio, ma ai fini della nostra analisi tali dati non sono necessari.

3.3.1 Analisi del dataset

A seguito della selezione degli attributi rilevanti al task condotto, viene effettuata un'analisi in merito alla distribuzione dei dati rispetto al genere, ai dialetti e alla combinazione delle due categorie.

Nella Figura 3.1 è riportata, a sinistra, la distribuzione dei dati rispetto al genere. Si osserva come il dataset risulti ben bilanciato, con una presenza pressoché equa di maschi e femmine. A destra, invece, è rappresentata la distribuzione degli esempi

in base ai dialetti considerati. Da tale grafico emerge un'evidente predominanza del dialetto Standard American English (SAE), che risulta significativamente più rappresentato rispetto agli altri. I restanti quattro dialetti (AAVE, Chicano English, Spanglish, e Other) mostrano invece una distribuzione relativamente bilanciata tra loro.

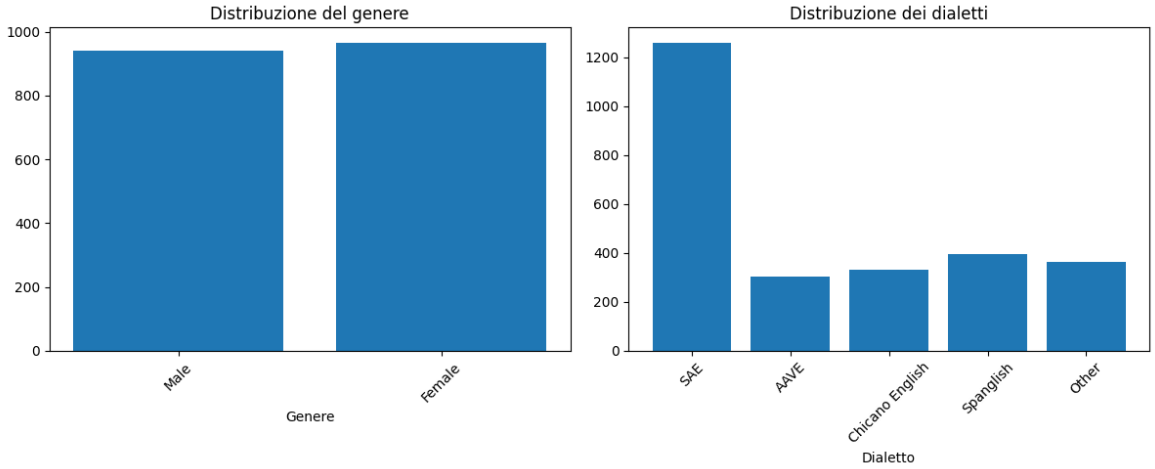

Figura 3.1: Distribuzione per Genere e Distribuzione per Dialetto

La Figura 3.2 mostra la distribuzione degli speaker in base al genere per ciascuno dei dialetti presenti nel dataset. Si osserva chiaramente una predominanza del dialetto SAE, con un numero particolarmente elevato di parlanti femminili rispetto a quelli maschili. Negli altri gruppi dialettali, la situazione è più variabile: AAVE, Chicano English e Spanglish presentano una marcata prevalenza di voci maschili, con una rappresentanza femminile piuttosto ridotta. Al contrario, la categoria Other risulta più equilibrata, con una leggera prevalenza di speaker donne. Questa distribuzione suggerisce uno sbilanciamento di genere all'interno di alcuni dialetti, un aspetto che potrebbe influenzare le performance dei modelli nei confronti dei diversi gruppi demografici.

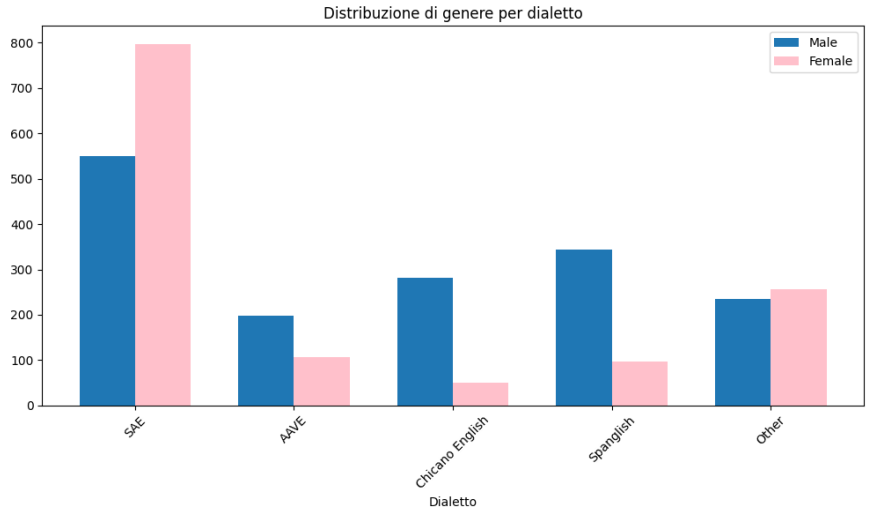

Figura 3.2: Distribuzione per Genere e Dialetto combinati

È da tener conto che per la selezione delle diverse categorie, l'attributo Gender mantiene l'esclusività dei valori mentre Dialect no. Questo sta a significare che, nel caso di Men e Women, consideriamo il caso in cui l'audio contenga esclusivamente solo uomini o solo donne, quindi non conversazioni comprendenti sia parlanti uomini che donne. Nel caso invece dell'attributo Dialect, nel considerare uno specifico dialetto, non andiamo ad escludere tutti gli altri. Quindi, ad esempio, se consideriamo gli audio contenenti spanglish (con spanglish == 1), non è detto che questi non contengano anche speakers appartenenti a dialetti differenti.

Per le successive fasi implementative condotte, il dataset è stato suddiviso secondo la proporzione 80-10-10: 80% per il train, 10% per la validation e 10% per il test, mantenendo invariata la distribuzione dei vari valori considerati per i vari attributi.

3.4 Implementazione LSTM e GRU

Un file audio può essere considerato una *Time Series*, in quanto rappresenta una sequenza ordinata di campioni audio acquisiti a intervalli di tempo regolari. Tuttavia, per essere utilizzato efficacemente in un task di *Automatic Speech Recognition (ASR)*, è necessario trasformare il segnale audio grezzo in una rappresentazione più informativa e compatibile con i modelli di deep learning.

Nel presente lavoro, sul dataset `spotify_podcast_ASR`, sono state implementate due architetture di Recurrent Neural Networks (RNN): **LSTM (Long Short-Term Memory)** e **GRU (Gated Recurrent Unit)**, al fine di modellare la sequenza temporale delle caratteristiche audio estratte.

Di seguito viene descritta la fase di preprocessing, che consiste nell'estrazione delle feature dal segnale audio, in modo da ottenere una serie temporale multivariata da fornire in input alle RNN.

3.4.1 Preparazione dei dati

Features audio

Al fine di rendere gli audio processabili da modelli di DL, in particolare da reti neurali ricorrenti (RNN), è stato necessario trasformare i segnali audio grezzi in una rappresentazione numerica più adatta alla modellazione sequenziale. In questo progetto, si è scelto di rappresentare ciascun file audio mediante il relativo spettrogramma, una forma di rappresentazione tempo-frequenza che cattura l'evoluzione spettrale del segnale nel tempo.

Questo processo consente di ottenere, per ogni audio, una matrice bidimensionale il cui asse verticale rappresenta le frequenze e quello orizzontale il tempo. Tali spettrogrammi vengono quindi utilizzati come input per i modelli RNN (LSTM e GRU), che possono così apprendere le caratteristiche temporali e acustiche del parlato in modo più efficace. In sintesi, il segnale audio grezzo viene trasformato in spettrogramma per evidenziare l'informazione acustica rilevante. Il risultato è una **Multivariate Time Series**, in cui a ciascun intervallo temporale è associato un vettore di feature che descrive l'energia o la potenza delle diverse bande di frequenza. In questo modo, lo spettrogramma consente di rappresentare l'audio non solo nel dominio del tempo, ma anche nel dominio della frequenza.

Features trascrizioni

Le trascrizioni, essendo in forma testuale, non sono processabili dalle RNN implementate. Devono quindi essere convertite in sequenze di numeri, per poter adeguatamente rappresentare le label dell'implementazione ed essere confrontabili con l'output generato dalla rete stessa. Per ottenere tali sequenze numeriche si effettua una fase di encoding in cui, a partire da un vocabolario predefinito contenente tutti i caratteri ammessi, si assegna a ciascun carattere un identificativo numerico univoco (id). Le trascrizioni vengono quindi convertite in sequenze di questi id, creando una rappresentazione numerica delle etichette testuali compatibile con i modelli neurali.

3.4.2 Definizione delle RNN

Poiché i dati ottenuti nella fase di preprocessing sono rappresentati come Multivariate Time Series, si è scelto di utilizzare due architetture di Recurrent Neural Networks (RNN) — in particolare, LSTM e GRU — per affrontare il task di Automatic Speech Recognition (ASR).

Queste reti sono progettate per elaborare sequenze nel tempo, e vengono impiegate per generare una sequenza di numeri che rappresentano identificativi univoci dei caratteri del testo parlato. Al seguito del processo di decodifica, questi identificativi numerici possono essere convertiti nella trascrizione testuale corrispondente all'audio fornito in input.

Nel presente lavoro si è scelto di implementare un modello di riconoscimento vocale *offline*, il quale ha accesso all'intera sequenza audio sin dall'inizio dell'elaborazione. In questo scenario, il modello riceve in input l'intera frase audio e produce in output una sequenza di identificativi numerici corrispondenti ai caratteri del testo trascritto.

Per sfruttare al meglio il contesto globale del segnale audio, si è optato per l'utilizzo di una **Bidirectional Recurrent Neural Network**. Questo tipo di architettura è particolarmente adatta ai modelli offline perché è in grado di analizzare le sequenze in entrambe le direzioni:

- in avanti (dal passato verso il futuro);
- all'indietro (dal futuro verso il passato).

Grazie a questa bidirezionalità, ogni frame di output della rete è calcolato tenendo conto sia delle informazioni precedenti che di quelle successive all'istante corrente. Questo approccio consente al modello di sfruttare l'intero contesto temporale durante la previsione, migliorando significativamente la precisione della trascrizione.

La Figura 3.3 illustra una rappresentazione schematica dell'architettura della RNN bidirezionale implementata.

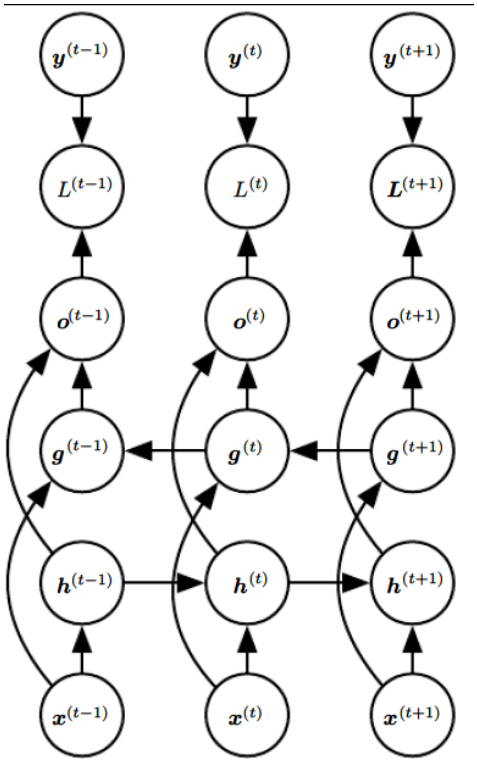

Figura 3.3: RNN Bidirezionale

Le RNN tradizionali presentano una difficoltà nel gestire le dipendenze a lungo termine all'interno delle sequenze. Questo limite è principalmente dovuto ai problemi del vanishing e exploding gradient durante il processo di apprendimento, che ostacolano l'aggiornamento efficace dei pesi nei passaggi molto distanti nel tempo.

Per affrontare questo problema, sono state utilizzate due varianti avanzate delle RNN: LSTM (Long Short-Term Memory) e GRU (Gated Recurrent Units). Entrambe queste architetture introducono meccanismi di gating interni, progettati per controllare il flusso delle informazioni attraverso la rete.

I gates decidono dinamicamente:

- quali informazioni mantenere nella memoria,
- quali aggiornare con nuovi dati,
- e quali scartare come non più rilevanti.

In questo modo, le LSTM e le GRU sono in grado di preservare informazioni utili nel tempo e di attenuare l'effetto dei gradienti che tendono a esplodere o a svanire, migliorando così la capacità del modello di apprendere relazioni anche a lungo termine all'interno della sequenza.

Architettura LSTM

La Figura 3.4 mostra l'architettura di una cella della LSTM. In particolare, ha una struttura interna che include una cell state (memoria) e tre gates (porte), che lavorano insieme per gestire le informazioni nel tempo.

Il compito principale della cell state è quello di trasportare le informazioni rilevanti lungo tutta la sequenza, funzionando da memoria per le predizioni future.

Le gates controllano il flusso delle informazioni, decidendo quali dati devono essere conservati e quali, invece, devono essere eliminati. Le tre gates sono:

- **Forget Gate**: decide quali informazioni eliminare dalla memoria;
- **Input Gate**: decide quali nuove informazioni aggiungere alla memoria;
- **Output Gate**: decide quali informazioni della memoria usare come output per il passo successivo.

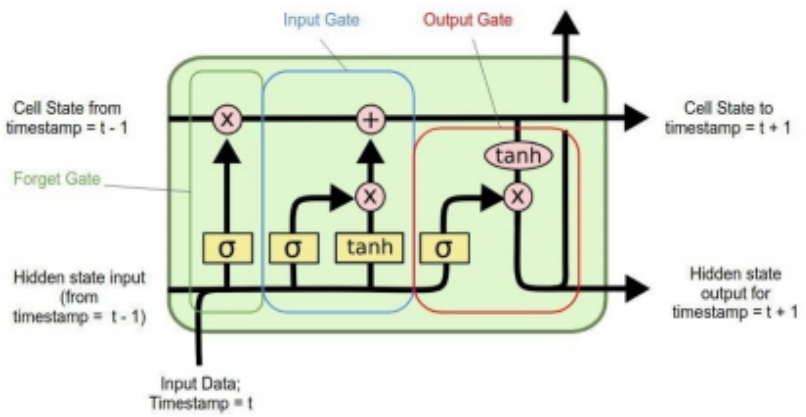

Figura 3.4: Architettura di una cella della LSTM

Architettura GRU

La Figura 3.5 mostra l'architettura di una cella della GRU. In particolare, ha una struttura interna che include due principali gate:

- **Update Gate**: Combinazione di due funzioni (dimenticare e aggiungere), che regola l'aggiornamento della memoria.
- **Reset Gate**: Elimina selettivamente informazioni passate, prevenendo l'esplosione del gradiente e focalizzandosi solo su ciò che è rilevante nel contesto attuale.

La GRU ha una struttura più semplice, rispetto dalla LSTM, ma altrettanto potente per gestire le sequenze temporali.

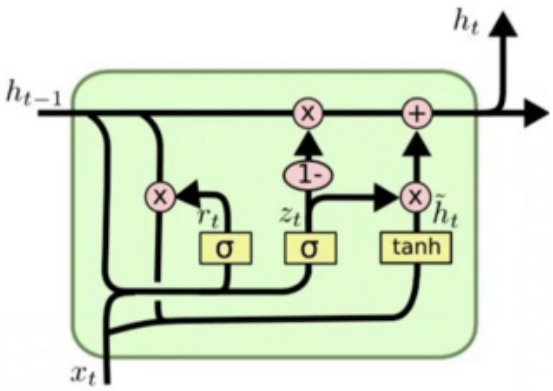

Figura 3.5: Architettura di una cella della GRU

Dense Layer

Dopo aver impilato più layer LSTM o GRU, sono stati aggiunti due layer densi finali. Questi layer hanno lo scopo di calcolare i valori grezzi (raw scores) associati a ciascun carattere nel vocabolario, invece di calcolare direttamente le probabilità. Questa scelta è stata fatta per sfruttare al meglio la funzione di perdita CTC (Connectionist Temporal Classification), come spiegato più dettagliatamente in seguito.

Scelta dei parametri

Per ottimizzare il modello LSTM o GRU, è importante trovare i migliori valori per gli iperparametri. Per fare ciò, è stata implementata la **Grid Search**, che esplora una combinazione di valori di parametri per identificare quelli che massimizzano le performance del modello. Il `param_grid` definisce questi iperparametri e i possibili valori da testare durante l'allenamento del modello.

```
param_grid = {
    'dropout_rate': [0.2, 0.5],
    'n_units': [64, 128],
    'n_layers': [1, 2, 3],
    'batch_size': [32, 64],
    'learning_rate': [0.001, 0.01]
}
```

Gli iperparametri sono:

- `dropout_rate`: Il dropout è una tecnica di regolarizzazione che aiuta a prevenire l'overfitting. Durante il train, una certa percentuale dei neuroni viene "spenta" casualmente ad ogni iterazione;
- `n_units`: Questo parametro definisce il numero di neuroni per ciascun layer LSTM o GRU. Un numero maggiore di unità può migliorare la capacità di apprendimento, ma potrebbe anche aumentare il rischio di overfitting e richiedere un training maggiore;

- `n_layers`: Questo parametro indica il numero di layer LSTM o GRU. Più layer possono aumentare la capacità di apprendimento della rete, ma a costo di maggiore complessità computazionale;
- `batch_size`: La batch size definisce quante istanze di dati vengono elaborate contemporaneamente durante la fase di training. Un valore maggiore riduce la varianza nelle stime del gradiente, ma potrebbe rallentare il train;
- `learning_rate`: Il learning rate controlla la velocità con cui i pesi della rete vengono aggiornati durante il training. Un valore troppo alto può portare a un apprendimento instabile, mentre un valore troppo basso può rallentare il processo di convergenza.

Il numero di epoche è impostato su 30, ma questo iperparametro non è incluso nella grid search, poiché viene gestito tramite il meccanismo di *EarlyStopping*. In particolare, abbiamo configurato l'EarlyStopping con una patience di 5 epoche e il monitoraggio del valore della metrica WER prodotta in fase di validazione.

Definizione delle Callback

Durante la fase di train del modello, sono state definite tre callback che monitorano specifici aspetti del processo di training e aiutano a migliorare le prestazioni e a risparmiare risorse computazionali. In particolare:

- `ModelCheckpoint`: lo scopo di tale callback è quello di salvare il modello ogni volta che viene trovato un miglioramento nella metrica Word Error Rate (WER) sulla validazione (ossia diminuisce);
- `EarlyStopping`: lo scopo di tale callbak è interrompere il train prima che venga raggiunto il numero massimo di epoche se il modello non mostra miglioramenti, e quindi modificare la combinazione di parametri. Questo avviene attraverso il monitoraggio della metrica WER in fase di validazione;
- `TensorBoard`: permette di monitorare il training in tempo reale attraverso i grafici dinamici di vari parametri.

Metriche WER e CER - Custom

Per la valutazione delle prestazioni dei modelli RNN sviluppati, sono state implementate due classi personalizzate: `CERMetric` e `WERMetric`, entrambe derivate dalla classe base `tensorflow.keras.metrics.Metric`. Queste classi sono specificamente progettate per calcolare la Character Error Rate (CER) e la Word Error Rate (WER), due metriche fondamentali per misurare l'accuratezza dei modelli di riconoscimento vocale automatico (ASR).

In particolare, ciascuna classe gestisce il post-processing dell'output prodotto dalle RNN, trasformando le sequenze di predizione (tipicamente come indici) in stringhe testuali confrontabili con le trascrizioni di riferimento. Il confronto viene poi effettuato tramite la libreria `JiWER`, che consente di calcolare automaticamente il numero di operazioni di sostituzione, inserimento e cancellazione necessarie per trasformare la sequenza predetta in quella corretta.

Queste metriche vengono quindi utilizzate durante la fase di validazione per monitorare l'effettiva qualità delle trascrizioni generate dal modello. Una descrizione formale e dettagliata delle metriche è disponibile nella Sezione 3.6.

Inoltre, è stato scelto di monitorare tramite callback la metrica Word Error Rate (WER) invece della Character Error Rate (CER) come principale indicatore di performance, al fine di garantire una valutazione coerente e direttamente confrontabile con i risultati riportati nel lavoro di Harris et al. [4]. Infatti, WER misura l'accuratezza delle trascrizioni a livello di parola — unità semantica più significativa rispetto ai singoli caratteri — per cui rappresenta una metrica più adatta per confronti a livello applicativo, soprattutto quando i benchmark di riferimento adottano lo stesso criterio di valutazione.

CTC Loss

La CTC Loss è una funzione di perdita sviluppata per affrontare compiti di allineamento temporale incerto, dove l'input e l'output non sono direttamente allineati. Questo problema si presenta frequentemente in task di ASR, dove l'input (il segnale

audio) e l'output (la trascrizione del parlato) hanno lunghezze diverse e non sono in sincrono.

In particolare, la CTC Loss consente di addestrare modelli di sequenze, come le RNN, senza la necessità di un allineamento esplicito tra ogni frame audio e il carattere corrispondente nella trascrizione. Il modello apprende così a generare output validi anche in presenza di allineamenti temporali variabili.

La funzione è stata introdotta da Graves et al. [5] ed è disponibile in diverse librerie di deep learning, tra cui TensorFlow, che ne fornisce un'implementazione tramite il metodo `tf.nn.ctc_loss`.

É stato definito il metodo `ctc_loss_fn`, una loss function personalizzata utilizzata per il calcolo della CTC Loss sui modelli RNN sviluppati. Questa funzione si occupa sia del preprocessing dei dati di input, adattandoli al formato richiesto dall'API `tf.nn.ctc_loss`, sia della chiamata diretta alla funzione di perdita. In particolare, prepara le sequenze di logit, le trascrizioni target e le rispettive lunghezze, garantendo la corretta compatibilità con i requisiti della `tf.nn.ctc_loss` in TensorFlow. Quest'ultima richiede in input un tensore contenente i logits temporali, ovvero i punteggi non normalizzati (raw scores) generati dall'ultimo livello della rete neurale.

Poiché `tf.nn.ctc_loss` applica internamente l'operazione di softmax lungo la dimensione delle label, l'output del modello non deve includere alcuna attivazione finale. Restituendo direttamente i logits grezzi, si garantisce il corretto calcolo della probabilità complessiva di tutte le possibili sequenze di allineamento tra l'input e la trascrizione, secondo il principio alla base della CTC Loss.

3.5 Inferenza con Whisper-medium Multilingual e Wav2Vec2-base

La seconda fase implementativa si concentra sull'utilizzo di modelli preaddestrati offerti da HuggingFace, al fine di effettuare un confronto con le precedenti strutture basate su RNN.

3.5.1 Modello Whisper-medium Multilingual

La scelta del modello openai/whisper-medium deriva dai risultati ottenuti all'interno dell'analisi condotta in [4], in cui viene effettuato un confronto approfondito tra diversi modelli preaddestrati disponibili su HuggingFace per il task di ASR. In particolare, il modello `openai/whisper-medium.en` si è distinto per le sue ottime prestazioni in termini di Word Error Rate (WER), risultando uno dei più accurati tra quelli considerati.

Tuttavia, a differenza di quanto fatto nello studio citato, in questo progetto si è scelto di utilizzare la versione multilingua del modello, ovvero `openai/whisper-medium`, anziché la variante ottimizzata esclusivamente per l'inglese. Questa decisione è motivata dalla natura del dataset utilizzato, che include anche segmenti di audio contenenti fenomeni di *code-switching*, in particolare tra inglese e spagnolo (come nel caso del dialetto Spanglish). L'uso della versione multilingua consente dunque una maggiore flessibilità e una migliore capacità di catturare tali alternanze linguistiche, offrendo un vantaggio nei contesti in cui il parlato non è interamente monolingue.

La versione selezionata del modello Whisper è la *medium*, che conta circa 769 milioni di parametri. Tale modello è stato addestrato su un ampio corpus costituito da circa 680.000 ore di dati audio annotati, utilizzando un approccio di supervisione debole su larga scala. Questo tipo di addestramento consente al modello di apprendere rappresentazioni robuste e generalizzabili, rendendolo particolarmente efficace anche in presenza di parlato spontaneo, rumore di fondo o variazioni linguistiche.

3.5.2 Modello Wav2Vec2-base

Il modello facebook/wav2vec2-base-960h è stato scelto per effettuare un'analisi comparativa con le performance dell'altro modello utilizzato.

Per questo progetto è stata adottata la versione *base* del modello Wav2Vec2, addestrata su 960 ore di dati audio tratti dal dataset LibriSpeech, campionati a 16 kHz. Per garantire quindi il corretto funzionamento del modello e l'ottenimento di risultati affidabili, è fondamentale che anche gli input audio forniti siano campionati alla stessa frequenza di 16 kHz. La versione utilizzata rappresenta una configurazione bilanciata tra complessità e prestazioni, risultando adatta per valutazioni comparative con architetture.

3.5.3 Inferenza dei modelli

A seguito della fase di selezione dei modelli, viene effettuata la fase di inferenza del modello per valutare le performance di questi ultimi in merito alle metriche Word Error Rate (WER) e Character Error Rate (CER) dei vari gruppi considerati. L'inferenza viene effettuata distintamente per Genere, Dialetto e la combinazione dei due attributi, al fine di individuare disparità in ASR in termini di diversi generi o dialetti considerati.

3.6 Metriche WER e CER

Nel contesto del riconoscimento vocale automatico (ASR), l'obiettivo principale è valutare l'accuratezza con cui il modello trascrive il parlato in testo. In questo lavoro, per tutti i modelli analizzati — sia quelli basati su RNN (LSTM e GRU), sia i modelli preaddestrati (Whisper-medium Multilingual e Wav2Vec2-base) — sono state calcolate le metriche sia a livello globale, sia suddividendo i risultati in base a genere e dialetto, oltre che considerando la combinazione di genere e dialetto.

Questa analisi dettagliata ha lo scopo di esaminare la fairness dei modelli, ovvero la loro capacità di fornire prestazioni uniformi e non discriminatorie rispetto a

caratteristiche socio-linguistiche degli speaker. Valutare l'impatto del genere e del dialetto consente di identificare eventuali bias nei modelli e di promuovere soluzioni più eque e inclusive.

In particolare le metriche analizzate sono: **Word Error Rate (WER)** e **Character Error Rate (CER)**. Queste metriche sono calcolate utilizzando la libreria `JiWER`.

Morris et al. [6] forniscono un confronto dettagliato delle metriche più comunemente utilizzate per valutare le prestazioni nei task di ASR.

3.6.1 Word Error Rate (WER)

Il WER è una misura che calcola la differenza tra la trascrizione corretta e quella prodotta dal sistema ASR. Viene calcolato come la somma del numero di inserimenti, sostituzioni ed eliminazioni necessarie per trasformare la trascrizione del modello in quella di riferimento.

$$WER = \frac{\text{Sostituzioni} + \text{Inserimenti} + \text{Eliminazioni}}{\text{Numero di parole}} \tag{3.6.0}$$

In particolare:

- *WER basso* indica che la trascrizione del modello è simile a quella corretta, quindi il modello ha buone performance;
- *WER alto* suggerisce che ci sono errori significativi nel riconoscimento delle parole.

3.6.2 Character Error Rate (CER)

Il CER misura invece l'errore a livello di caratteri, ed è simile al WER, ma invece di considerare le parole, si confrontano i singoli caratteri.

$$CER = \frac{\text{Sostituzioni} + \text{Inserimenti} + \text{Eliminazioni}}{\text{Numero di caratteri}} \tag{3.6.0}$$

Come nel caso del WER, un CER elevato indica la presenza di errori significativi, mentre un CER basso suggerisce che il numero di errori è ridotto.

In particolare, la metrica WER offre un'analisi complessiva della trascrizione a livello di parole, che è spesso l'obiettivo principale in molte applicazioni di riconoscimento vocale. La metrica CER è invece utile per una valutazione più fine a livello di caratteri, consentendo l'analisi ortografica delle parole prodotte.

3.7 Pipeline Metodologia

La Figura 3.6 mostra l'intero workflow per il task di Automatic Speech Recognition, illustrando nel dettaglio le fasi fondamentali della pipeline implementata. Tale processo è stato progettato per essere modulare e flessibile, in modo da poter integrare diverse strategie di preprocessing e tipologie di modelli.

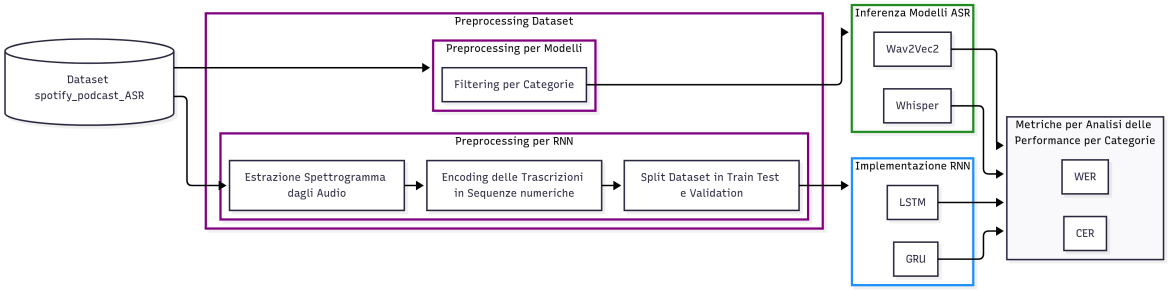

Figura 3.6: Pipeline Metodologia

CAPITOLO 4

Analisi dei Risultati

All'interno di una pipeline di Deep Learning, una delle fasi più rilevanti è rappresentata dalla valutazione e analisi dei risultati, in quanto consente di misurare l'efficacia dei modelli e di trarre considerazioni utili in relazione agli obiettivi di ricerca. In questo capitolo si propone un'analisi dettagliata delle performance di due modelli preaddestrati ampiamente utilizzati per il riconoscimento vocale automatico e disponibili tramite la piattaforma HuggingFace: Wav2Vec2 e Whisper.

I modelli sono stati testati seguendo la stessa suddivisione introdotta nei capitoli precedenti, al fine di mantenere coerenza metodologica nell'analisi. In particolare, i risultati sono stati valutati sull'intero dataset e poi analizzati in maniera disaggregata per genere, dialetto e per la combinazione delle due categorie. Questo approccio permette di evidenziare eventuali disparità nelle performance dei modelli in relazione a caratteristiche socio-linguistiche, e di comprendere meglio i loro punti di forza e le loro limitazioni.

Va sottolineato che l'analisi si concentra esclusivamente sui modelli preaddestrati. Sebbene inizialmente fosse previsto il confronto anche con modelli RNN addestrati da zero, la fase di training di tali modelli è stata temporaneamente sospesa a

causa dell'elevato costo computazionale richiesto. L'esecuzione completa di questi esperimenti è stata pertanto rimandata a sviluppi futuri.

L'obiettivo dell'analisi è duplice: da un lato, valutare la capacità dei modelli di generalizzare su un dataset che presenta una notevole varietà linguistica; dall'altro, rispondere in maniera sistematica alle domande di ricerca delineate nella Sezione 3.2.

4.1 RQ1: Quale architettura di rete neurale dei modelli preaddestrati garantisce le migliori prestazioni in termini di WER e CER sull'intero dataset?

RQ$_1$. *Quale architettura di rete neurale dei modelli preaddestrati — Whisper-medium multilingual e Wav2Vec2-base — garantisce le migliori prestazioni in termini di Word Error Rate (WER) e Character Error Rate (CER) sull'intero dataset?*

Per rispondere alla prima domanda di ricerca, sono state analizzate le distribuzioni complessive dei valori di WER e CER relativi ai campioni elaborati da entrambi i modelli preaddestrati.

4.1.1 Whisper-medium multilingual

Le Figure 4.1 e 4.2 illustrano le prestazioni del modello Whisper-medium multilingual rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

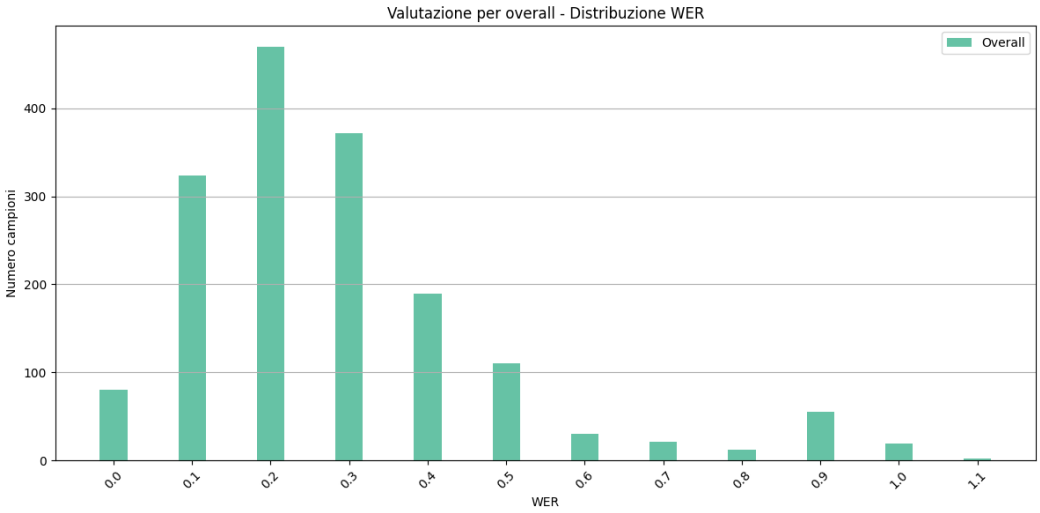

Figura 4.1: Distribuzione dei campioni in base al tasso di errore WER per Whisper-medium multilingual

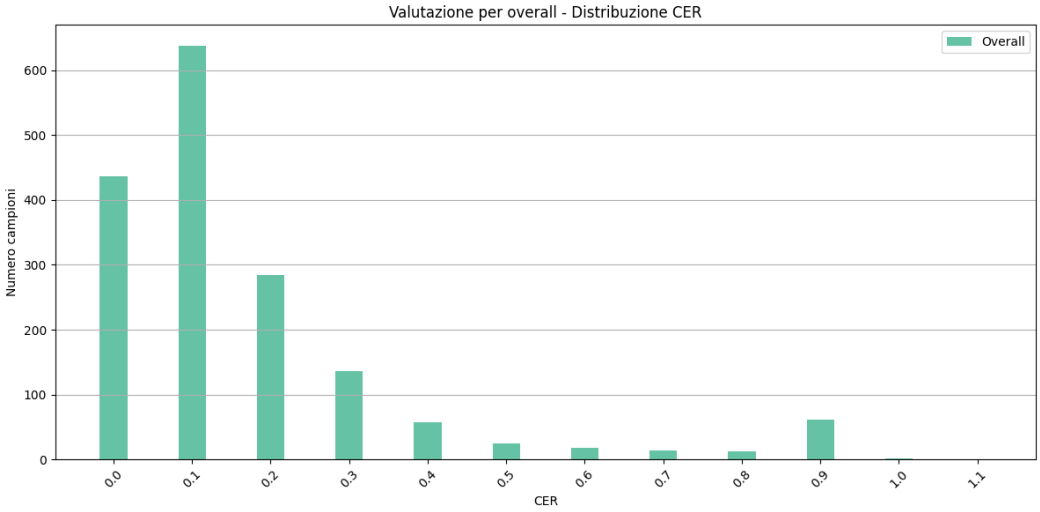

Figura 4.2: Distribuzione dei campioni in base al tasso di errore CER per Whisper-medium multilingual

Dall'analisi della Figura 4.2 emerge che la maggior parte dei campioni presenta valori di CER molto bassi, segno di un numero limitato di errori a livello di carattere. La distribuzione è infatti concentrata nei range di errore più contenuti, indicando un'elevata precisione nel riconoscimento dei singoli caratteri.

Per quanto riguarda la metrica WER, si osserva una concentrazione prevalente dei campioni nell'intervallo $[0.1, 0.4]$, suggerendo un numero di errori per parola

superiore rispetto a quelli riscontrati a livello di carattere.

In entrambi i casi, le distribuzioni mostrano un rapido calo della frequenza all'aumentare del tasso di errore, con pochi campioni che registrano valori elevati di WER o CER. Questo comportamento complessivo indica che il modello Whisper-medium multilingual è in grado di generalizzare efficacemente sul dataset, garantendo buone prestazioni sia nel riconoscimento delle parole che dei caratteri.

Pertanto, si può concludere che il comportamento complessivo del modello Whisper-medium multilingual risulta particolarmente positivo per entrambe le metriche, dimostrando affidabilità sia nel riconoscimento delle parole che dei singoli caratteri.

4.1.2 Wav2Vec2-base

Le Figure 4.3 e 4.4 illustrano le prestazioni del modello Wav2Vec2-base rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

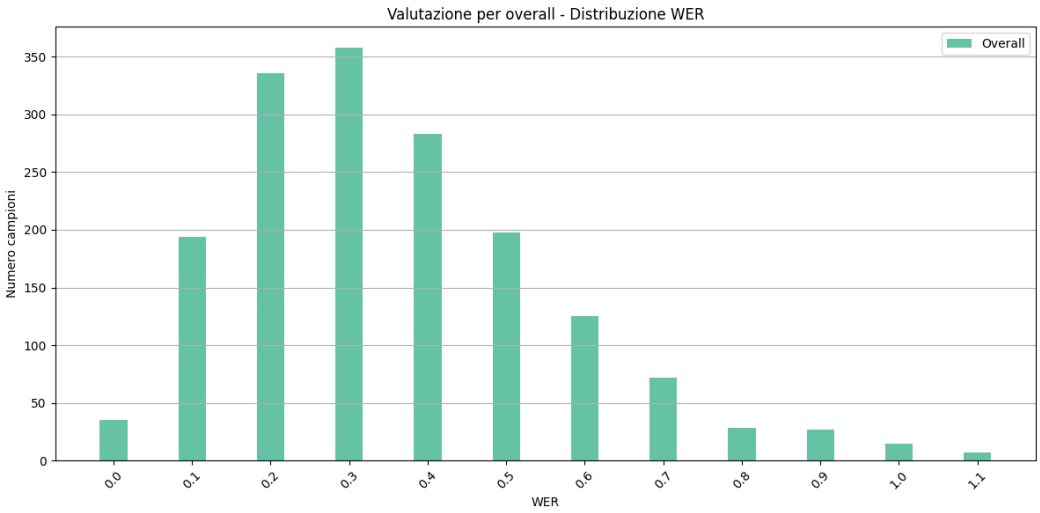

Figura 4.3: Distribuzione dei campioni in base al tasso di errore WER per Wav2Vec2-base

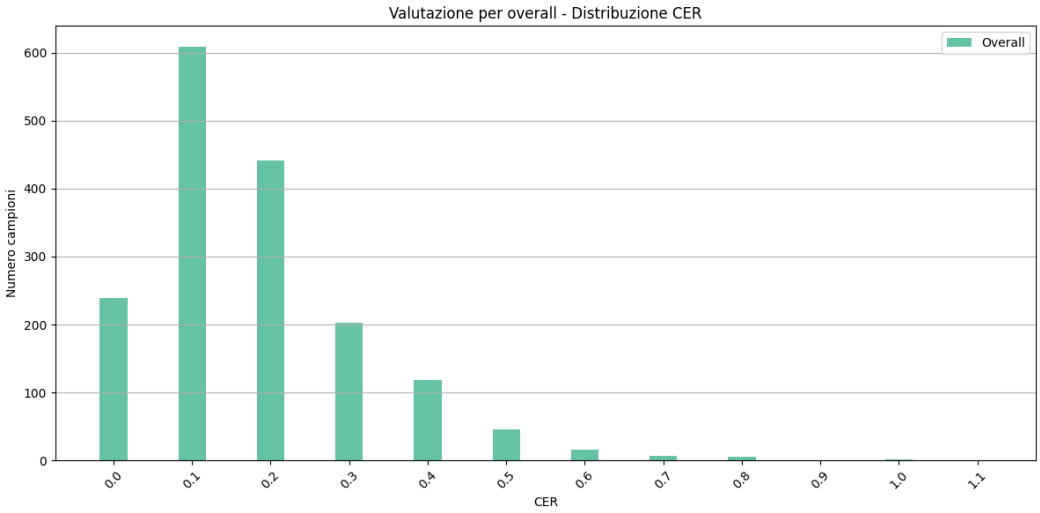

Figura 4.4: Distribuzione dei campioni in base al tasso di errore CER per Wav2Vec2-base

Analogamente a quanto osservato per il modello Whisper-medium multilingual, anche nel caso del Wav2Vec2-base la metrica CER evidenzia prestazioni migliori rispetto alla WER. Questo conferma un comportamento comune nei sistemi di riconoscimento vocale: è generalmente più semplice riconoscere correttamente i singoli caratteri piuttosto che l'intera parola, poiché il riconoscimento lessicale richiede una maggiore comprensione del contesto linguistico.

Tuttavia, analizzando la distribuzione dei valori di WER per Wav2Vec2-base, si nota che la maggior parte dei campioni si concentra in un intervallo più ampio, compreso tra $[0.1, 0.5]$, rispetto a quello osservato per Whisper. Questa maggiore dispersione nei range di errore indica una maggiore variabilità nelle prestazioni e una tendenza a commettere più errori nel riconoscimento delle parole.

4.1.3 Confronto dei modelli

Nel complesso, il confronto tra i due modelli evidenzia che Wav2Vec2-base mostra prestazioni inferiori, soprattutto nel riconoscimento delle parole, risultando meno preciso rispetto a Whisper-medium multilingual sull'intero dataset. Questo vantaggio di Whisper è attribuibile alla sua capacità multilingue, che gli consente di gestire con maggiore flessibilità la varietà linguistica presente nei dati, migliorando così l'accuratezza delle predizioni.

👉 **Answer to RQ_1.** Il modello preaddestrato che ha complessivamente le prestazioni migliori - sia per la metrica WER sia per la metrica CER - è Whisper-medium multilingual.

4.2 RQ2: In quali gruppi sociolinguistici si osservano i più alti tassi di errore, suggerendo la presenza di bias sistematici nei modelli ASR?

🔍 **RQ_2.** *In quali gruppi sociolinguistici (definiti da dialetto e genere separatamente) si osservano i più alti tassi di errore, suggerendo la presenza di bias sistematici nei modelli ASR?*

Per rispondere alla seconda domanda di ricerca, sono state analizzate in modo separato le distribuzioni dei valori di WER (Word Error Rate) e CER (Character Error Rate) in funzione del genere e del dialetto dei parlanti, considerando i campioni elaborati da entrambi i modelli preaddestrati.

Questa analisi ha l'obiettivo di individuare eventuali disparità sistematiche nelle prestazioni dei modelli ASR, verificando se alcuni gruppi sociolinguistici mostrano tassi di errore significativamente più elevati.

4.2.1 Whisper-medium multilingual

Distribuzione per dialetto

Le Figure 4.5 e 4.6 illustrano le prestazioni distinte per tipologia di dialetto del modello Whisper-medium multilingual rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

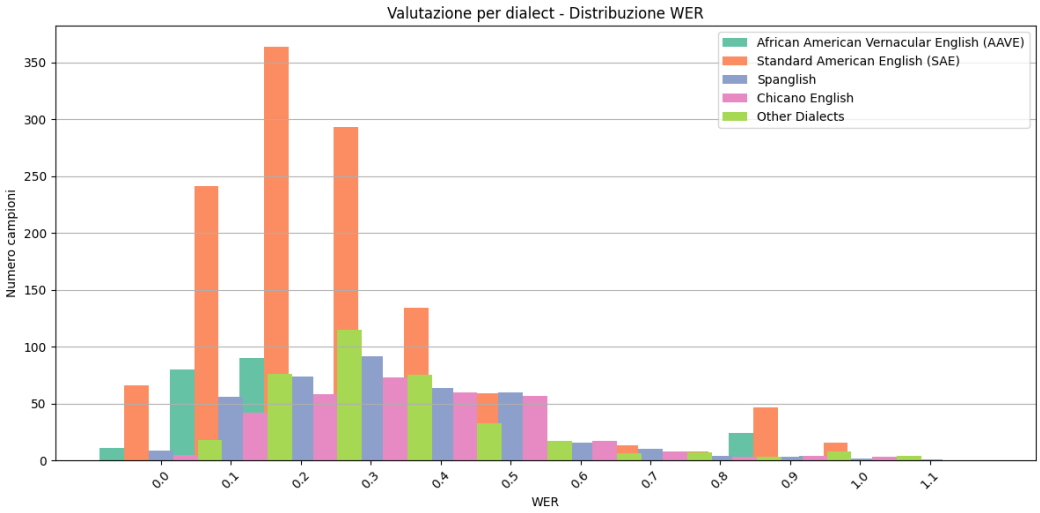

Figura 4.5: Distribuzione dei campioni per dialetto in relazione al tasso di errore WER nel modello Whisper-medium multilingual

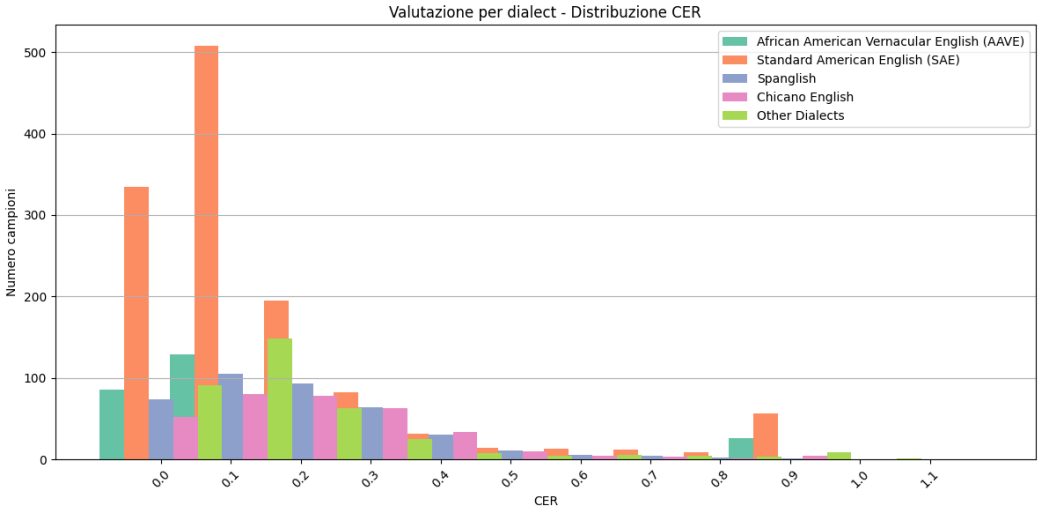

Figura 4.6: Distribuzione dei campioni per dialetto in relazione al tasso di errore CER nel modello Whisper-medium multilingual

Il dialetto SAE (Standard American English), essendo il più rappresentato nel dataset, mostra naturalmente picchi più elevati nella distribuzione. Tuttavia, non è solo la quantità di campioni a distinguerlo: i valori di errore associati a questo dialetto sono anche significativamente più bassi.

Nello specifico, per la metrica WER, il dialetto SAE mostra una distribuzione compresa principalmente nell'intervallo $[0.0, 0.4]$, mentre gli altri dialetti tendono

a estendersi su un range più ampio e più elevato, $[0.0, 0.6]$. Allo stesso modo, per la metrica CER, i campioni SAE si concentrano maggiormente nel range $[0.0, 0.2]$, mentre i dialetti rimanenti si distribuiscono fino a $[0.3]$.

Queste differenze indicano che Whisper-medium multilingual riconosce il parlato in dialetto SAE con maggiore precisione, mentre tendono a commettere più errori con gli altri dialetti. Ciò suggerisce la presenza di un bias sistematico a favore del dialetto SAE, con prestazioni sensibilmente inferiori sui parlanti che utilizzano varianti dialettali differenti.

Distribuzione per genere

Le Figure 4.5 e 4.6 illustrano le prestazioni distinte per gernere del modello Whisper-medium multilingual rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

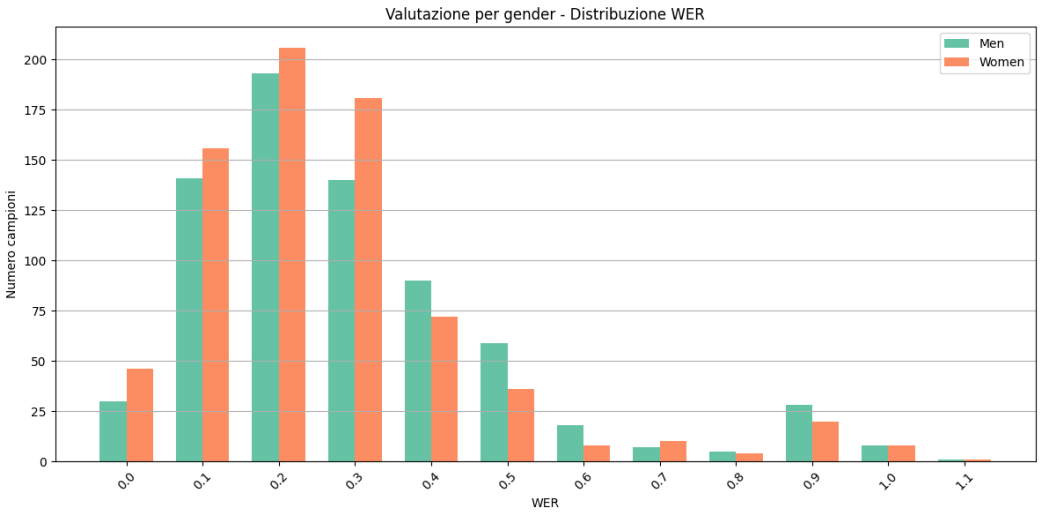

Figura 4.7: Distribuzione dei campioni per genere in relazione al tasso di errore WER nel modello Whisper-medium multilingual

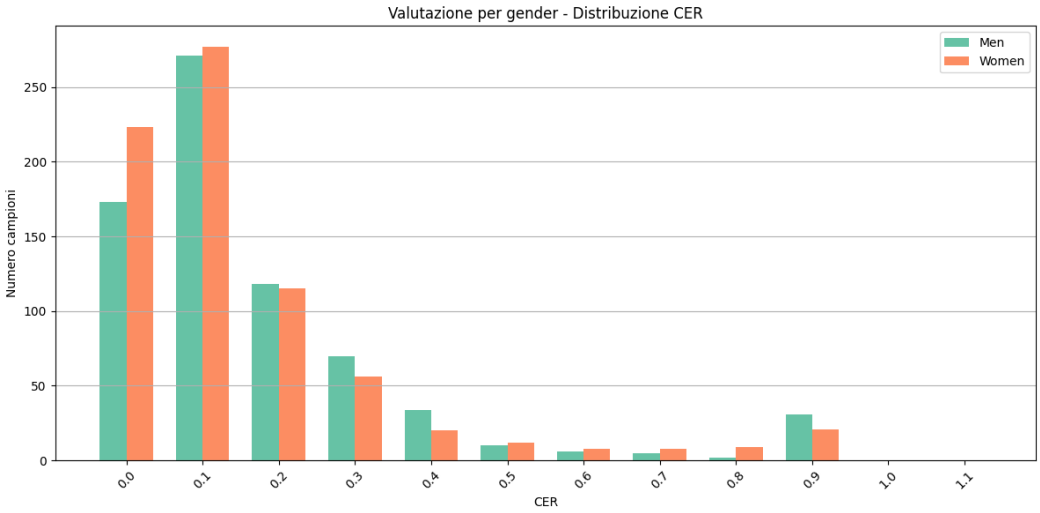

Figura 4.8: Distribuzione dei campioni per genere in relazione al tasso di errore CER nel modello Whisper-medium multilingual

Dall'analisi dei grafici emerge che le prestazioni del modello Whisper-medium multilingual sono nel complesso equilibrate tra i generi, sia per quanto riguarda la metrica WER (Word Error Rate) sia per la CER (Character Error Rate). Le distribuzioni per genere non mostrano differenze marcate, ad eccezione di alcuni picchi isolati.

Tuttavia, osservando più attentamente le fasce di errore più basse — in particolare WER < 0.4 e CER < 0.2 — si nota che il numero di campioni femminili è maggiore rispetto a quelli maschili. Questo suggerisce che, pur mantenendo prestazioni simili tra i due generi in termini generali, il modello tende a riconoscere con maggiore precisione le voci femminili, ottenendo un numero più consistente di trascrizioni corrette o con pochi errori.

Pertanto, si può ipotizzare una leggera preferenza del modello verso il parlato femminile, anche se il bilanciamento tra i generi rimane sostanzialmente stabile.

4.2.2 Wav2Vec2-base

Distribuzione per dialetto

Le Figure 4.9 e 4.10 illustrano le prestazioni distinte per tipologia di dialetto del modello Wav2Vec2-base rispetto alle metriche WER (Word Error Rate) e CER

(Character Error Rate).

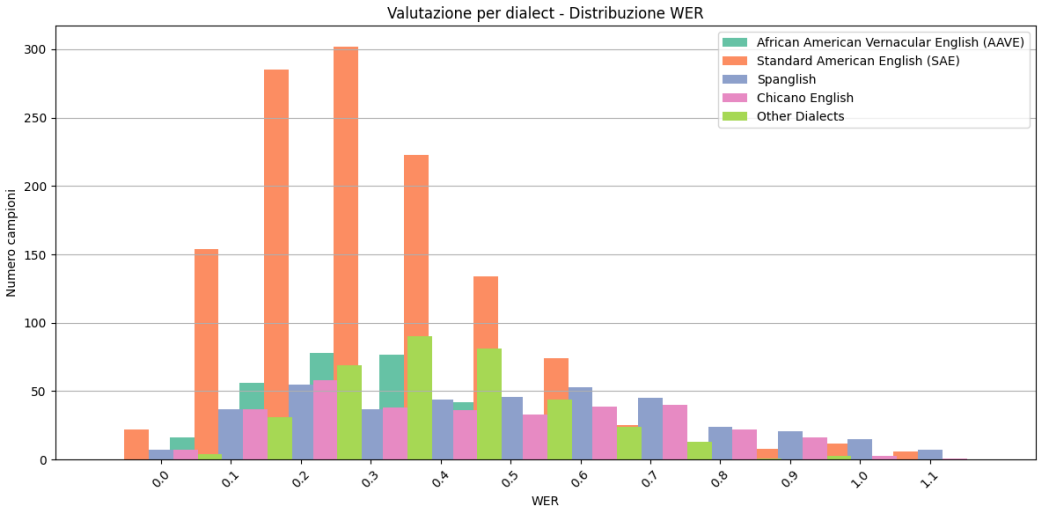

Figura 4.9: Distribuzione dei campioni per dialetto in relazione al tasso di errore WER nel modello Wav2Vec2-base

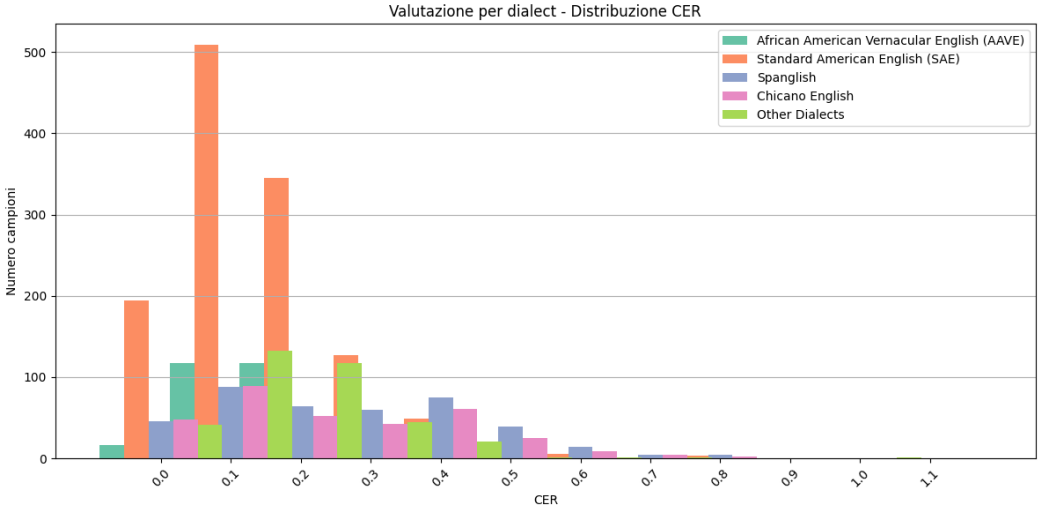

Figura 4.10: Distribuzione dei campioni per dialetto in relazione al tasso di errore WER nel modello Wav2Vec2-base

Per quanto riguarda la metrica WER, il dialetto SAE (Standard American English) mostra una distribuzione degli errori prevalentemente concentrata nell'intervallo $[0.0, 0.5]$, mentre gli altri dialetti si estendono su un range più ampio e tendenzialmente più elevato, fino a $[0.8]$. Analogamente, per la metrica CER, i campioni relativi

al dialetto SAE si concentrano principalmente nel range $[0.0, 0.2]$, mentre per gli altri dialetti la distribuzione raggiunge valori anche fino a $[0.5]$.

Queste differenze indicano che il modello Wav2Vec2-base è in grado di riconoscere il parlato in dialetto SAE con maggiore accuratezza, mentre presenta prestazioni più deboli nel riconoscimento dei dialetti alternativi. Questo comportamento evidenzia la presenza di un bias sistematico a favore del dialetto SAE, a discapito dei parlanti che utilizzano altre varianti linguistiche.

Inoltre, rispetto al modello Whisper-medium multilingual, le distribuzioni degli errori del Wav2Vec2-base risultano più ampie, sia per la WER che per la CER. Ciò suggerisce una maggiore variabilità nelle prestazioni e una minore capacità di generalizzazione del modello rispetto alle differenze dialettali presenti nel dataset.

Distribuzione per genere

Le Figure 4.11 e 4.12 illustrano le prestazioni distinte per genere del modello Wav2Vec2-base rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

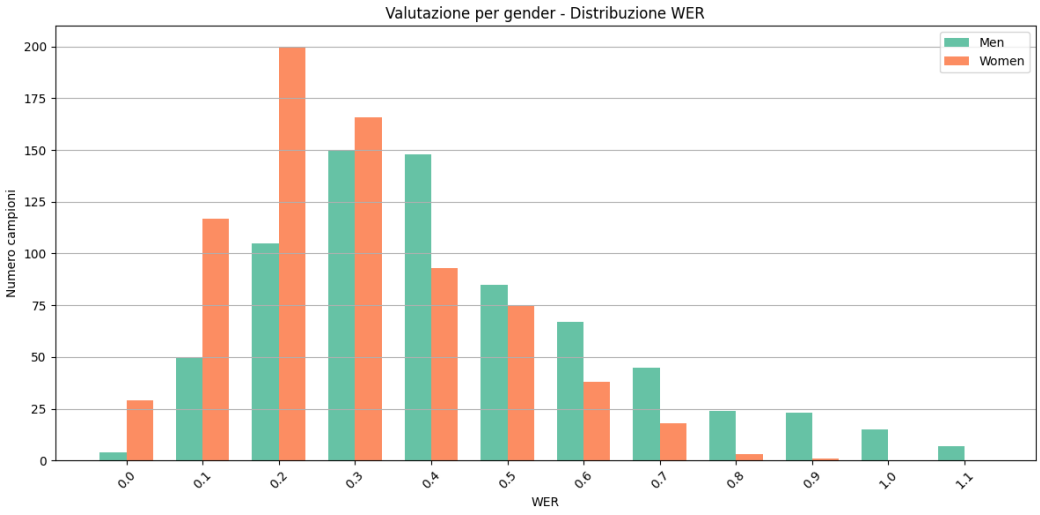

Figura 4.11: Distribuzione dei campioni per genere in relazione al tasso di errore WER nel modello Wav2Vec2-base

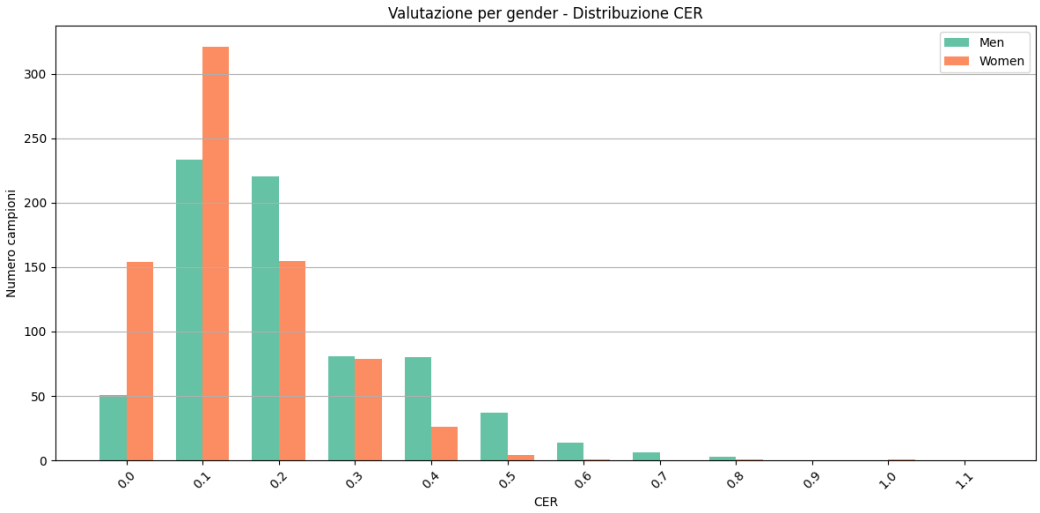

Figura 4.12: Distribuzione dei campioni per genere in relazione al tasso di errore CER nel modello Wav2Vec2-base

L'analisi dei risultati mostra che il modello Wav2Vec2-base presenta un bias di genere più evidente rispetto al modello Whisper-medium multilingual. In particolare, si osserva che il numero di campioni femminili con bassi tassi di errore — ossia WER < 0.4 e CER < 0.2 — è significativamente superiore rispetto ai campioni maschili.

Questa disparità indica che il modello Wav2Vec2-base è più preciso nel riconoscere voci femminili, mentre tende a commettere un numero maggiore di errori nel caso di voci maschili. A differenza di Whisper, che mostra una distribuzione più bilanciata tra i generi, Wav2Vec2-base manifesta un gender bias più marcato, penalizzando in modo sistematico un genere rispetto all'altro.

Tale comportamento potrebbe essere attribuito a uno sbilanciamento nel dataset di addestramento o a una mancanza di generalizzazione del modello rispetto alle caratteristiche vocali maschili, e rappresenta un aspetto critico da considerare nell'impiego di sistemi ASR in contesti reali ed equi.

4.2.3 Confronto dei modelli

L'analisi delle distribuzioni degli errori in funzione del dialetto e del genere conferma che, in generale, la metrica CER (Character Error Rate) registra tassi di errore

più contenuti rispetto alla WER (Word Error Rate), con una maggiore concentrazione di campioni nei range di errore più bassi. Questo indica che i modelli tendono ad essere più precisi nel riconoscimento dei singoli caratteri rispetto alle parole intere, comportamento comune nei sistemi ASR.

Confrontando i due modelli considerati, emerge chiaramente che Wav2Vec2-base è quello che manifesta il bias più marcato, sia dal punto di vista del genere che del dialetto.

Il modello Whisper-medium multilingual, pur mostrando una certa variabilità nei risultati in base al dialetto, mantiene una distribuzione pressoché bilanciata tra i generi, con prestazioni simili per voci maschili e femminili. Le principali differenze, in questo caso, si osservano tra dialetti, con una preferenza per il dialetto più rappresentato (SAE).

Al contrario, Wav2Vec2-base evidenzia disparità significative sia tra i generi che tra i dialetti. Le voci femminili tendono a ottenere risultati nettamente migliori rispetto a quelle maschili, e il modello mostra una forte dipendenza dal dialetto SAE, con prestazioni sensibilmente peggiori sugli altri dialetti.

Questi risultati suggeriscono che Wav2Vec2-base è più soggetto a bias sistematici, penalizzando in modo rilevante alcuni gruppi sociolinguistici. Di conseguenza, il modello Whisper-medium multilingual risulta nel complesso più equo e robusto, soprattutto in termini di equità tra generi.

👉 **Answer to RQ_2.** Il modello Wav2Vec2-base evidenzia un bias sistematico sia verso dialetti (a meno del SAE) che verso il genere, con prestazioni significativamente peggiori per voci maschili e dialetti meno rappresentati. Al contrario, Whisper-medium multilingual mostra una distribuzione più bilanciata, con un bias principalmente legato al dialetto.

4.3 RQ3: Quale dialetto mostra il più basso gender bias in termini di performance, in modo ricorrente su tutti i modelli ASR valutati?

RQ$_3$. *Quale dialetto mostra il più basso gender bias in termini di performance, in modo ricorrente su tutti i modelli ASR valutati?*

4.3.1 Whisper-medium multilingual

Le Figure 4.13 e 4.14 illustrano le prestazioni distinte per tipologia di dialetto e genere del modello Whisper-medium multilingual rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

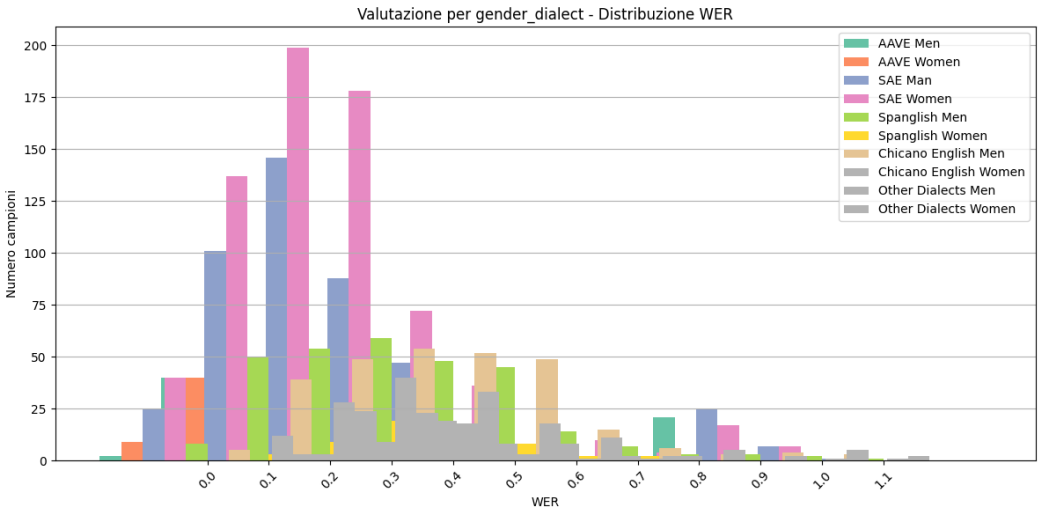

Figura 4.13: Distribuzione dei campioni per genere e dialetto in relazione al tasso di errore WER nel modello Whisper-medium multilingual

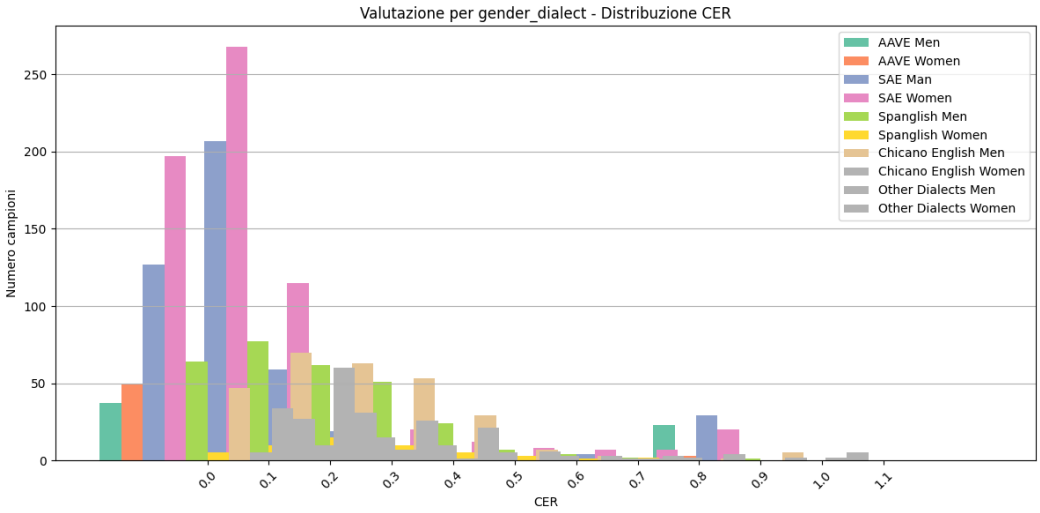

Figura 4.14: Distribuzione dei campioni per genere e dialetto in relazione al tasso di errore CER nel modello Whisper-medium multilingual

Il modello Whisper-medium multilingual mostra differenze significative tra i generi solo nel caso del dialetto SAE. In particolare, per questo dialetto, si osserva una chiara disparità di prestazioni a favore del genere femminile. Al contrario, per tutti gli altri dialetti, le prestazioni risultano omogenee tra voci maschili e femminili, suggerendo un comportamento più equo del modello rispetto al genere in contesti dialettali diversi dal SAE.

4.3.2 Wav2Vec2-base

Le Figure 4.15 e 4.16 illustrano le prestazioni distinte per tipologia di dialetto e genere del modello Wav2Vec2-base rispetto alle metriche WER (Word Error Rate) e CER (Character Error Rate).

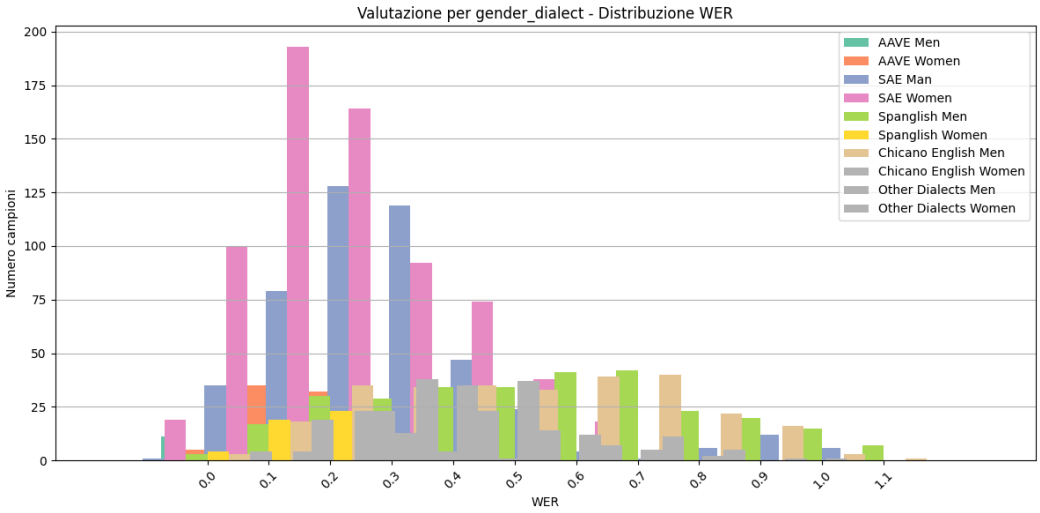

Figura 4.15: Distribuzione dei campioni per genere e dialetto in relazione al tasso di errore WER nel modello Wav2Vec2-base

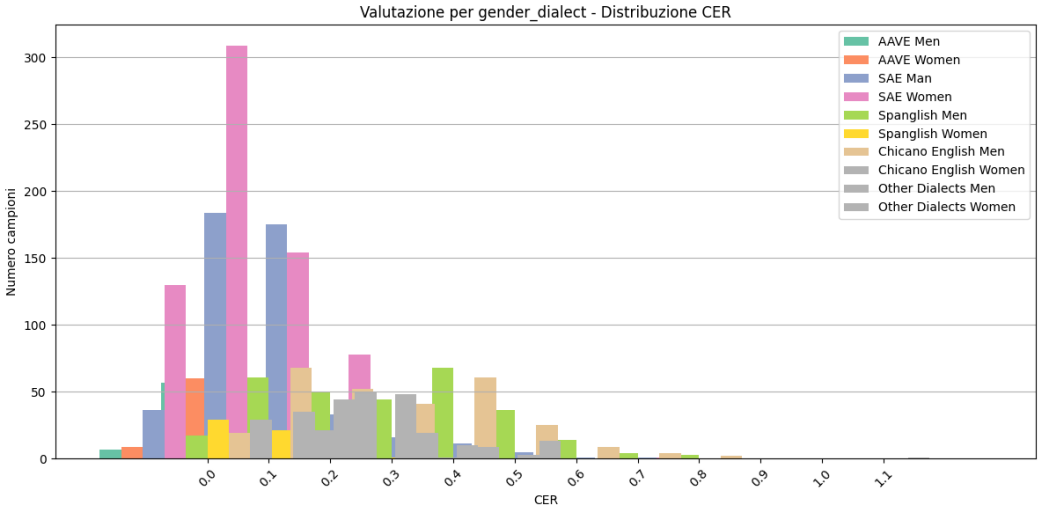

Figura 4.16: Distribuzione dei campioni per genere e dialetto in relazione al tasso di errore CER nel modello Wav2Vec2-base

Anche il modello Wav2Vec2-base presenta un comportamento analogo a quello osservato per Whisper-medium multilingual: il dialetto SAE evidenzia una marcata disparità di prestazioni tra genere maschile e femminile, con risultati significativamente migliori per le voci femminili. Non sono presenti particolari differenze tra gli altri dialetti.

4.3.3 Confronto dei modelli

Nel complesso, entrambi i modelli mostrano prestazioni comparabili sia in termini di WER (Word Error Rate) che di CER (Character Error Rate) nell'analisi congiunta di genere e dialetto.

👉 **Answer to RQ$_3$.** Il dialetto che presenta il più alto gender bias per entrambi i modelli è il SAE, poiché mostra prestazioni significativamente migliori per il genere femminile. Per gli altri dialetti, invece, le prestazioni tendono a rimanere simili indipendentemente dal genere.

CAPITOLO 5

Conclusioni e Sviluppi Futuri

5.1 Conclusioni

L'analisi comparativa tra i modelli Whisper-medium multilingual e Wav2Vec2-base, condotta su un dataset diversificato per dialetti e generi, ha permesso di evidenziare importanti differenze in termini di prestazioni e di equità.

In primo luogo, il modello Whisper-medium multilingual si è dimostrato il più efficace sull'intero dataset, ottenendo i valori più bassi di Word Error Rate (WER) e Character Error Rate (CER). Ciò conferma la maggiore capacità di generalizzazione di questa architettura, probabilmente dovuta alla sua struttura multilingue e alla maggiore robustezza nel trattamento della variabilità linguistica (RQ1).

Per quanto riguarda l'equità tra gruppi sociolinguistici (RQ2), entrambi i modelli mostrano alcune forme di bias sistematici. Tuttavia, Wav2Vec2-base risulta più sensibile a queste variazioni, con prestazioni peggiori per i parlanti di dialetti diversi dal SAE e per il genere maschile. Al contrario, Whisper-medium multilingual presenta un comportamento più bilanciato rispetto al genere, ma evidenzia comunque differenze di prestazione tra i dialetti, suggerendo la presenza di un bias dialettale residuo.

Infine, l'analisi congiunta di dialetto e genere (RQ3) ha mostrato che il dialetto SAE è quello che presenta il più alto gender bias per entrambi i modelli, con prestazioni significativamente migliori per le voci femminili. Gli altri dialetti, invece, mostrano prestazioni più stabili tra i generi, suggerendo che il gender bias sia fortemente legato a questa particolare variante linguistica.

Nel complesso, i risultati evidenziano la necessità di sviluppare modelli ASR più equi e inclusivi, capaci di offrire prestazioni costanti anche in presenza di varietà sociolinguistiche. Sebbene Whisper-medium multilingual si sia dimostrato più performante e meno soggetto a bias rispetto a Wav2Vec2-base, persistono criticità legate soprattutto alla gestione dei dialetti, che meritano ulteriori approfondimenti in contesti applicativi e di ricerca futura.

5.2 Sviluppi Futuri

Il progetto presenta diverse possibili direzioni di sviluppo, tra cui:

- Eseguire il fine-tuning dei modelli preaddestrati su dati bilanciati per dialetto, al fine di migliorare l'equità delle prestazioni e ridurre i bias sistematici. In particolare, allenare i modelli su campioni rappresentativi di più varianti dialettali può contribuire a sviluppare sistemi ASR più fairness-oriented, capaci di garantire risultati omogenei e affidabili indipendentemente dalla varietà linguistica del parlato.
- L'esecuzione dei modelli RNN implementati, al momento non effettuata a causa dell'elevato costo computazionale richiesto.
- Le RNN implementate sono attualmente di tipo bidirezionale, il che consente un'analisi offline dell'audio. Una possibile evoluzione consiste nell'implementare una RNN unidirezionale, che permetterebbe di effettuare predizioni in tempo reale (streaming ASR), rendendo il modello utilizzabile per applicazioni online con bassa latenza.
- La pipeline sviluppata, che integra sia modelli RNN sia modelli preaddestrati, potrebbe essere applicata a dataset in lingua italiana. Questo rappresenta un

contributo significativo, dato che le analisi di fairness sui dialetti italiani sono ancora poco esplorate in letteratura. In tale direzione, sarebbe opportuno valutare le soluzioni proposte finora tenendo conto dei dataset in italiano presentati nel survey di Giordano et al. [7].

- Si potrebbero testare ulteriori modelli preaddestrati e confrontarne le prestazioni in termini di WER e CER, seguendo un approccio simile a quello adottato da Harris et al. [4], al fine di identificare le architetture più efficaci per specifici contesti o varianti linguistiche.

- Un'ulteriore linea di ricerca potrebbe riguardare l'analisi di metriche alternative o complementari a WER e CER, come proposto nel lavoro di Morris et al. [6], per ottenere una valutazione più completa delle performance dei modelli ASR sviluppati.

Bibliografia

[1] I.-E. Veliche, Z. Huang, V. A. Kochaniyan, F. Peng, O. Kalinli, and M. L. Seltzer, "Towards measuring fairness in speech recognition: Fair-speech dataset," *arXiv preprint arXiv:2408.12734*, 2024. (Citato a pagina 5)

[2] M. Xia, A. Field, and Y. Tsvetkov, "Demoting racial bias in hate speech detection," *arXiv preprint arXiv:2005.12246*, 2020. (Citato a pagina 5)

[3] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6162–6166. (Citato a pagina 5)

[4] C. Harris, C. Mgbahurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 15 166–15 184. (Citato alle pagine 6, 9, 20, 22 e 45)

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376. (Citato a pagina 21)

[6] A. C. Morris, V. Maier, and P. D. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition." in *Interspeech*, 2004, pp. 2765–2768. (Citato alle pagine 24 e 45)

[7] M. Giordano and C. Rinaldi, "A survey on spoken italian datasets and corpora," *IEEE Access*, 2025. (Citato a pagina 45)