Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

Implementazione di un sistema visivo di riconoscimento basato sul labiale con integrazione dell'informazione della lingua parlata

Relatori
Prof. Andrea Francesco Abate
Dott.ssa Lucia Cascone

Candidato
Maurizio Ricco
Matricola: 0512108622

Anno Accademico 2021/2022

Non stancarti di correre verso il futuro.

Indice

Abstract

Con l'ascesa dei sistemi informatici fa sempre comodo proteggere i propri dati all'interno con l'ausilio di parole chiavi, le cosiddette *password*, però con l'evoluzione delle tecnologie informatiche stanno diventando poco sicuri e possono essere "*facilmente*" reperibili. Si è pensato così ad introdurre l'*identificazione biometrica* che consente di utilizzare una password biometrica che rappresenta un tratto fisiologico e/o comportamentale. Nello specifico, nel campo dell'identificazione biometrica abbondano gli studi dove affrontano tale problematica soprattutto utilizzando la zona labiale, *focus* principale di questo lavoro. Nella seguente trattazione, oltre all'identificazione biometrica tramite le labbra, verrà considerata il riconoscimento della lingua parlata per riuscire ad ottenere un miglioramento nell'identificazione. Si percorrerà la strada del Deep Learning implementando un modello apposita tramite la libreria *Keras* cercando di ottimizzare al meglio lo spazio di memoria che necessita quest'ultimo per un funzionamento corretto. La particolarità di tale scelta è la comodità di lavorare direttamente con sequenze visive dove quest'ultime verranno convertite in dati binari per renderle comprensibili al modello in fase di *Preprocessing*. Il dataset utilizzato è stato elaborato totalmente per questo lavoro, comprendendo 8 lingue: *Italiano, Inglese, Tedesco, Spagnolo, Olandese, Russo, Giapponese e Francese*.

Ognuno delle lingue trattate si è voluto avere lo stesso numero di soggetti ottenendo così un dataset *bilanciato*. La *cross-validation* del dataset sarà 80% training - 20% testing e le sperimentazioni affrontate saranno principalmente due che consentiranno di raggiungere l'obiettivo prefissato per questo lavoro: *"Riconoscimento della lingua parlata"* e *"Identificazione del soggetto"*. Corrispettivamente si otterranno il *74,21%* e *49,20%* di accuratezza, e in quest'ultima con l'integrazione della lingua parlata si otterrà il *6%* di miglioramento nell'identificazione biometrica tramite la zona labiale.

Capitolo 1

Introduzione

Il seguente lavoro ha come obiettivo lo sviluppo di algoritmi che utilizzano modelli neurali ricorrenti per effettuare l'identificazione di una persona con l'integrazione del riconoscimento della lingua parlata tramite la zona labiale e l'analisi del movimento labiale del soggetto parlante. Come verrà reso evidente nel corso del lavoro, le strategie utilizzate per lo studio e la risoluzione del problema si basano sul Deep Learning. Questa scelta è caratterizzata in quanto la rete neurale ricorrente viene allenata utilizzando direttamente le sequenze costituite dai frame che compongono i video dei soggetti parlanti. Nel Capitolo 2, verranno confrontati vari lavori già svolti da altri ricercatori, in modo tale da far comprendere al lettore come possono essere sfruttate le informazioni riguardo alle labbra. Successivamente saranno approfondite le implementazioni delle strategie e delle metodologie utilizzate nel Capitolo 3, si prosegue con il Capitolo 4 dove verranno descritti i diversi risultati dati dalle sperimentazioni. Concludendo nel Capitolo 5 saranno presenti le conclusioni e possibili sviluppi futuri.

1.1 Introduzione del problema

Negli ultimi tempi, si può notare che su un qualunque dispositivo che si ha in possesso, esso è capace di identificare il suo legittimo proprietario tramite ad un corrispettivo tratto biometrico. Tale processo viene chiamata *identificazione biometrica* e con l'evoluzione della tecnologia offre la possibilità di identificare il proprietario del dispositivo tramite, ad esempio, la sua faccia oppure dal suono della sua voce. Ci sono stati moltissimi e profondi lavori riguardo all'identificazione biometrica utilizzando tratti visivi e/o uditivi che oggi si possono usufruire. Altri, invece, sono ancora in fase di progresso che puntano all'obiettivo di utilizzare altri tratti biometrici in modo tale da avere più informazioni nell'identificazione di una persona. In questi lavori, ancora in corso d'opera, i ricercatori si focalizzano sul tratto visivo della zona labiale e che da questo tratto si possono trarre molte informazioni riguardo al soggetto. Si stanno adottando molte strategie per permettere tale identificazione, tra queste è utilizzo del fattore sia visivo che uditivo per aver un'accuratezza sufficiente, oppure, approcci più sintetici, ovvero eliminare informazioni che possono essere inutili, come l'audio, dove tale scelta è dovuta dal fatto di alleggerire il carico di dati e concentrarsi solo sulla forma e sul movimento delle labbra (ad esempio, in approcci Learning-based). Quindi da come si può intuire, diverse ricerche puntano ad ottenere un'accuratezza elevata per l'identificazione biometrica tramite la zona labiale, nominato *stato dell'arte*. Una soluzione per aumentare l'identificazione, che verrà descritta nel corso di questo lavoro, si basa principalmente sull'identificazione biometrica tramite la zona labiale con

l'integrazione della lingua parlata. L'idea è di aggiungere un informazione in più ad un soggetto, ovvero la sua lingua madre, e capire se questa informazione comporta ad un miglioramento nell'identificazione senza avere la necessità di utilizzare l'audio. Quindi, in questo lavoro, verranno descritte due sperimentazioni, che verranno successivamente messe insieme, saranno sviluppate allo stesso modo e con lo stesso modello neurale. L'integrazione della lingua parlata potrebbe essere un'informazione molto interessante dal fatto che il linguaggio è da sempre una delle abilità distintive dell'essere umano, e per questo ci sono molti lavori che sfidano l'*Intelligenza Artificiale* ad ottenere la lingua parlata da un soggetto percorrendo sia la strategia di rimozione dell'audio che l'ausilio di esso.

1.2 Sistema di riconoscimento biometrico

Un sistema di riconoscimento biometrico è un particolare tipo di sistema informatico che ha la funzionalità e lo scopo di identificare una persona sulla base di una o più caratteristiche fisiologiche e/o comportamentali, confrontandole con i dati precedentemente acquisiti e presenti nel database del sistema, tramite degli algoritmi e di sensori di acquisizione dei dati in input. La biometria è la scienza che studia le grandezze biofisiche allo scopo di identificare tratti fisici e comportamentali di un essere umano. Durante il corso del tempo, sono stati identificati fattori da utilizzare che permettono di valutare l'idoneità di qualsiasi tratto da utilizzare nell'autenticazione biometrica.

- **Universalità:** ogni persona che usa quel sistema biometrico deve possedere quel tratto;
- **Unicità:** il tratto dovrebbe essere sufficientemente differente per ogni individuo in modo tale da renderlo unico e distinguibile dagli altri;
- **Permanenza:** legata alla maniera in cui il tratto evolve nel tempo;
- **Collezionabilità:** la velocità nell'acquisizione o nella misurazione del tratto.

I sistemi biometrici sono caratterizzati da un processo di utilizzo che, in linea di massima, si può ricondurre ad una operazione di confronto di una caratteristica fisica o comportamentale acquisita da un soggetto con uno o più campioni della stessa precedentemente registrati. Sia la registrazione che il confronto vengono realizzati secondo la successione di passi raffigurati in Figura 1.1.

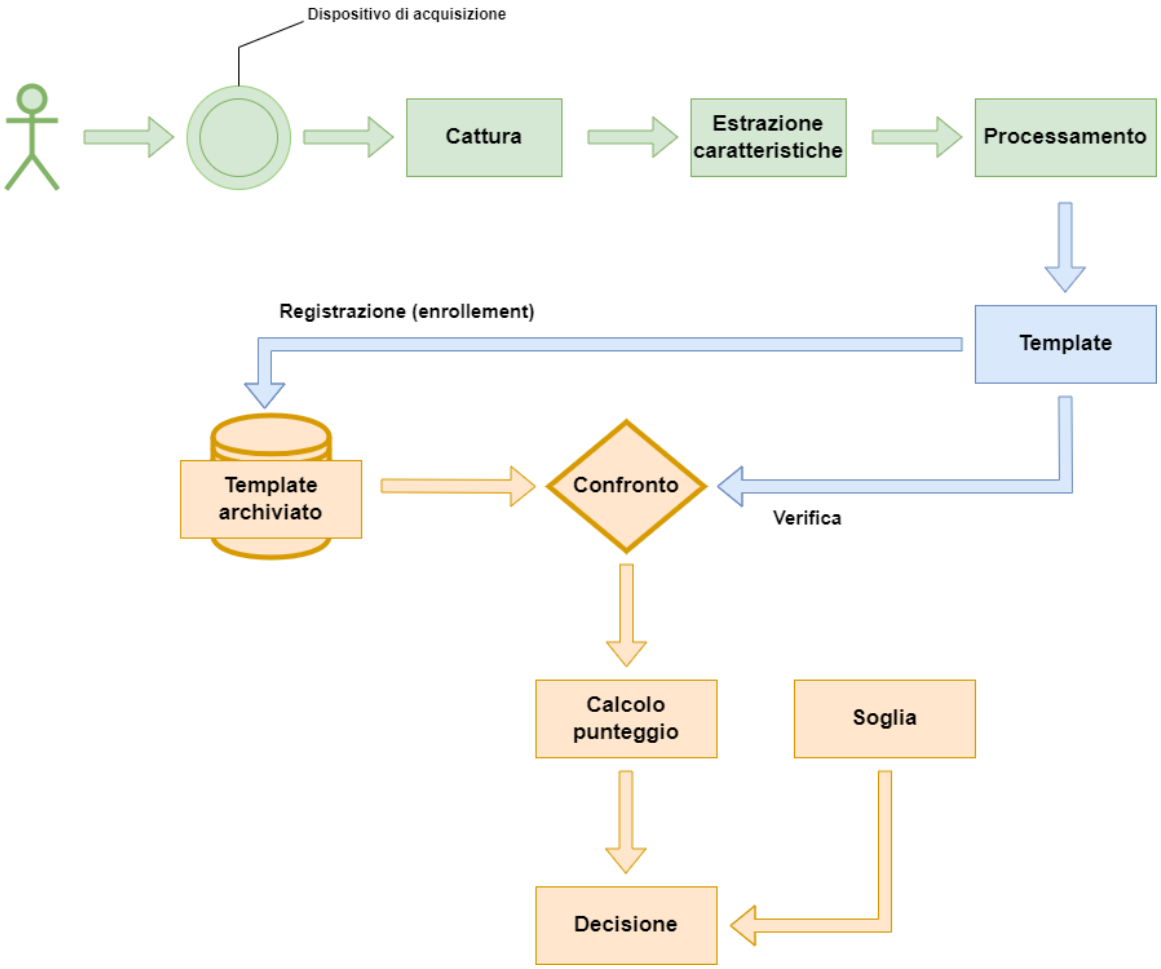

Figura 1.1: Processo biometrico

Un processo di un sistema biometrico è diviso nelle seguenti fasi:

1. **Enrollment**
2. **Verifica**
3. **Autenticazione**
4. **Riconoscimento/Autenticazione**

Nel processo di registrazione (chiamato anche *"Enrollment"*), l'utente fornisce al sistema biometrico una sua caratteristica fisica o comportamentale per mezzo di un dispositivo di acquisizione (ad esempio uno scanner per impronte digitali o una videocamera). Il campione viene processato per estrarne le informazioni caratteristiche distintive, che formano il cosiddetto *template* che si può definire come una rappresentazione matematica dei dati biometrici. Il *template* viene quindi memorizzato nel sistema in modo tale da essere utilizzato come confronto durante la fase di autenticazione. I sistemi biometrici possono operare in due diverse modalità: verifica e identificazione. Il processo di **verifica** (match 1-a-1) si ha quando il soggetto dichiarare la sua identità. Il sistema quindi effettua un confronto tra il tratto biometrico del soggetto rilevato in tempo reale e quella corrispondente del *template* presente nell'archivio. L'**identificazione** (match 1-a-molti) si ha quando il tratto biometrico acquisito in tempo reale viene confrontata con tutti gli altri tratti presenti nel database del sistema e viene poi associato al tratto più simile. Per aumentare la sicurezza del sistema di riconoscimento si possono fondere più tecniche biometriche grazie ai sistemi *multimodali*. Questi permettono un riconoscimento più accurato e diminuiscono il *failure-to-enroll rate*, ovvero il tasso di errore.

Ci sono due tipologie di errore:

- **FRR** (False Rejection Rate) è la percentuale di falsi rifiuti, utenti autorizzati ma respinti per errore, in pratica il sistema non riesce a riconoscere le persone autorizzate.
- **FAR** (False Acceptance Rate) è la percentuale di false accettazioni, utenti non autorizzati ma accettati per errore, il sistema quindi accetta le persone che non sono autorizzate.

Per regolare il rapporto tra i falsi rifiuti e le false accettazioni, si può definire la variabile t come il grado di tolleranza del sistema. Se questo grado è basso si ha un numero elevato di false accettazioni, con un grado alto invece si ha un numero elevato di falsi rifiuti. Tramite le funzioni si può calcolare l'EER (*Equal Error Rate*) $FAR(t*) = FRR(t*) = EERt*$ rappresenta il punto di equilibrio del sistema attraverso il quale è possibile regolare il rapporto FRR/FAR. Nelle applicazioni reali i valori di tolleranza si trovano al di sotto di $t*$ per garantire un numero ridotto di false accettazioni.

1.3 Breve riassunto sull'Intelligenza Artificiale

La maggior parte dei lavori che hanno permesso di sviluppare sistemi biometrici si basano sull'Intelligenza Artificiale, dove è quel ramo dell'informatica che consente la programmazione e progettazione dei sistemi informatici di dotare alle macchine di specifiche caratteristiche che vengono considerate tipicamente umane, come ad esempio percezioni visive e decisionali. L'Intelligenza Artificiale nasce con l'ascesa dei computer e si riconosce la sua nascita all'interno degli anni '50, tramite il programma Logic Theorist sviluppato da due ricercatori informatici, Allen Newell e Herbert Simon, era infatti in grado di dimostrare alcuni teoremi di matematica partendo da determinate informazioni. Ebbe così inizio di un nuovo percorso nella storia dell'informatica dove moltissime università e aziende informatiche, tra cui in particolare l'IBM, puntarono alla ricerca e allo sviluppo di nuovi programmi e software in grado di pensare e agire come gli essere umani almeno in determinati contesti. L'Intelligenza Artificiale non fu solo una svolta per gli informatici, ma consentì di semplificare il lavoro in determinati campi, come ad esempio nel 1969, dove alcuni studenti e ricercatori del Carnegie Institute of Technology realizzarono un programma, denominato **DENDRAL**, che era in grado di ricostruire una molecola semplice a partire dalle informazioni ottenute dallo spettrometro di massa. La nuova era dell'Intelligenza Artificiale si apre con il nuovo utilizzo di un algoritmo che, già ideato alla fine degli anni '60, non aveva trovato la massima applicazione. Si tratta dell'algoritmo che consentiva l'apprendimento per reti neurali, le cui sperimentazioni coprirono sia ambienti informatici e sia quelli psicologici.

1.3.1 Machine Learning

Uno dei principali successi nella storia dell'Intelligenza Artificiale è stata fatta quando si sono potuti ricreare degli algoritmi specifici, in grado di far migliorare il comportamento della macchina (nel senso capacità nel saper scegliere le giuste decisioni ed di agire nel modo corretto) così tali algoritmi sono capaci di imparare tramite l'esperienza, cercando proprio di simulare gli esseri umani. Lo sviluppo di algoritmi, in grado di imparare dai propri errori, è un concetto fondamentale per realizzare sistemi intelligenti che operano in contesti per i quali i programmatori non possono prevedere tutte le possibilità di sviluppo e i contesti in cui il sistema si trova a operare. Tramite l'apprendimento automatico (**Machine Learning**), una macchina è in grado di imparare a svolgere una determinata azione anche se tale azione non è mai stata programmata tra quelle possibili. La complessità dell'apprendimento automatica ha portato a dover suddividere due differenti possibilità, a seconda delle richieste di apprendimento che vengono fatte alla macchina. Si parla di **apprendimento supervisionato** e di **apprendimento non supervisionato**. La differenza tra queste due modalità sta soprattutto nel differente contesto entro cui si deve muovere la macchina per apprendere le regole generali e particolari che lo portano alla conoscenza. Nell'*apprendimento supervisionato* la macchina ha a disposizione come esempi degli obbiettivi che deve raggiungere, mostrando le relazioni tra input, output atteso e risultato, dove quest'ultimo viene dato dalla macchina. Dall'insieme dei dati mostrati, la macchina deve essere in grado di ottenere un metodo che le permette di scegliere il corretto output per il raggiungimento dell'obiettivo. Nel caso dell'*apprendimento non supervisionato*, invece, la macchina dovrà essere in grado di effettuare scelte senza aver visto niente riguardo alle differenti possibilità di output a seconda degli input selezionati. In questo caso il computer impara dai propri errori. Successivamente, l'apprendimento automatico ha reso possibile lo sviluppo delle reti neurali artificiali, ossia un particolare modello matematico che, ispirandosi ai neuroni e alle reti neurali degli umani, punta alla soluzione di diversi problemi a seconda della conoscenza degli input e i risultati ottenuti a seconda delle scelte effettuate. Dal punto di visto matematico,

una rete neurale può essere definita come una funzione composta, ossia dipendente da altre funzione a loro volta definibili in maniera diversa a seconda di ulteriori funzioni dalle quali esse dipendono.

1.3.2 Deep Learning

Il Deep Learning, conosciuto anche come **apprendimento profondo**, è una sottocategoria del Machine Learning e indica quel ramo dell'intelligenza artificiale che fa riferimento agli algoritmi ispirati alla struttura e alla funzione del cervello umano, chiamati reti neurali artificiali. Il Deep Learning fa parte di un insieme di metodi di Machine Learning basati sull'assimilazione di rappresentazioni di dati, al contrario degli algoritmi per l'esecuzione di compiti specifici. Le strutture di Deep Learning sono, per esempio, state applicati nella computer vision, ovvero nel riconoscimento automatico della lingua parlata, nell'elaborazione della lingua naturale, audio e nella bioinformatica. Si può definire il Deep Learning 1.2 come un sistema che prende in considerazione una classe di algoritmi di apprendimento automatico che:

- usano diversi livelli di unità non lineari a cascata per compiere compiti di estrazione di caratteristiche e di trasformazione; Ciascun livello in considerazione utilizza l'uscita del livello procedente come input;
- sono basati sull'apprendimento non supervisionato di livelli gerarchici multipli di caratteristiche (e di rappresentazioni) dei dati;
- apprendono multipli livelli di rappresentazione che corrispondono a differenti livelli di astrazione; questi livelli formano una gerarchia di concetti.

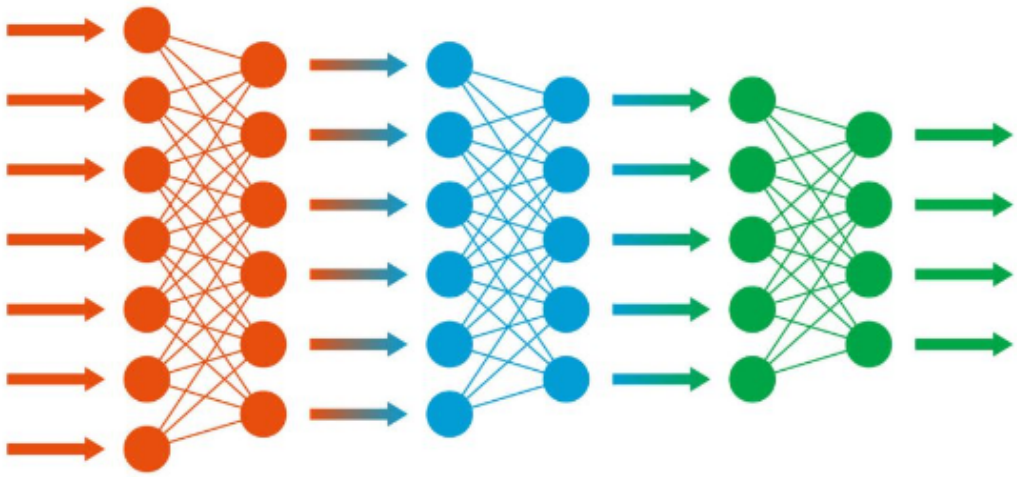

Figura 1.2: Struttura gerarchica dell'apprendimento Deep Learning

Nel Capitolo 3, verranno approfonditi al livello teorico il funzionamento del Deep Learning e il concetto di rete neurale.

1.4 Riconoscimento facciale

Un esempio di sistema di riconoscimento biometrico è il riconoscimento facciale, in cui si definisce una tecnica di intelligenza artificiale utilizzata in biometria per identificare o verificare l'identità di una persona a partire da una o più immagini che la ritraggono. Il suo funzionamento avviene mediante tecniche di elaborazione digitale delle immagini, ignorando tutto quello che non rappresenta una faccia, come edifici, alberi, corpi e altro, dove vengono solitamente definiti *background*. Si può affermare che si tratta di un riconoscimento di pattern, dove il pattern da riconoscere è il viso umano. Per facilitare il riconoscimento di una faccia, i primi sistemi progettati tenevano conto che un viso umano è composto da due occhi, un naso e una bocca. I sistemi di oggi invece riescono a riconoscere una persona anche se quest'ultima ha il viso ruotato, o comunque non in visione frontale. Tale riconoscimento può avvenire modellando la faccia come un oggetto da due dimensioni (2D) oppure da tre dimensioni (3D). Tra i principali algoritmi utilizzati per il riconoscimento facciale troviamo:

- PCA (Principal Component Analysis): permette di ottenere da uno spazio ad alta dimensionalità (pari al numero di pixel dell'immagine) un sottospazio significativo ai fini del riconoscimento.
- LDA (Linear Discriminant Analysis): oltre a ottenere un sottospazio di dimensionalità minore, come nel caso della PCA, permette una suddivisione in classi all'interno delle quali la varianza è minima.
- Metodi Kernel.
- Modello di Markov nascosto.

Il riconoscimento facciale, genericamente, può essere utilizzato in sistemi real-time, ovvero in brevissimo tempo la persona che si trova davanti ad una fotocamera di un dispositivo che fornisce questa funzionalità esso viene identificato.

1.5 Regione di interesse (ROI)

Una regione di interesse (ROI) è un sottoinsieme di un'immagine o di un set di dati identificato per uno scopo preciso. Il set di dati può essere uno dei seguenti:

- *Set di dati forma d'onda o 1D:* il ROI è un intervallo di tempo o frequenza sulla forma d'onda (un grafico di una quantità tracciata rispetto al tempo).
- *Set di dati immagini o 2D:* il ROI è definito da determinati limiti su un'immagine di un oggetto o su un disegno.
- *Volume o set di dati 3D:* il ROI sono i contorni o le superfici che definiscono un oggetto fisico

. In breve, la regione di interesse, nel campo biometrico, è la zona per la quale aiuta ad un sistema di riconoscimento biometrico ad identificare o verificare un individuo oppure uno specifico set di dati. Ad esempio un sistema biometrico facciale, la regione di interesse è il viso di un individuo; un sistema biometrico basato sulla zona labiale sono le labbra 1.3, ecc. Questo aiuta a capire che in molti lavori tramite l'intelligenza artificiale noi abbiamo bisogno che il modello venga addestrato cercando di dargli in input dati che possano servire nello scopo che si è posti, motivo per cui il modello non ha bisogno di addestrarsi su elementi che non gli servono e che appesantirebbe di molto la sua computazione.

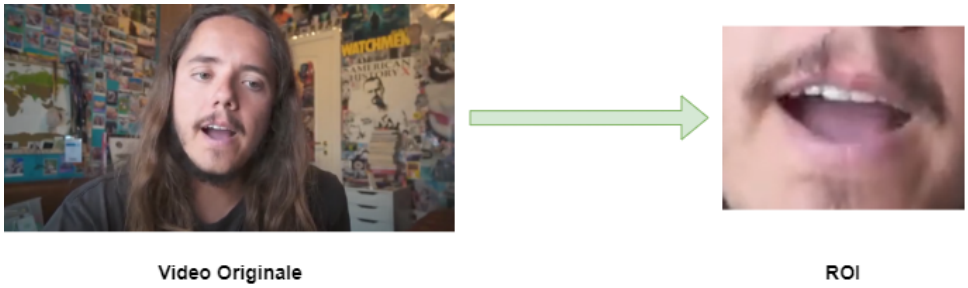

Figura 1.3: Esempio di ROI

1.6 Situazione di oggi

Nella società odierna, i dispositivi mobili come telefoni e laptop sono considerati essenziali sia per scopi personali che lavorativi e i rischi delle password usati come unico mezzo di autenticazione sono un enorme problema. Nonostante sia in continuo sviluppo efficacia dei sistemi informatici, soprattutto in potenza di calcolo, anche i metodi per scovare parole chiavi, che permettono di autenticare un utente, siano pericolosi. Quindi, è essenziale avere un'autenticazione sicura prima di accedere ai dispositivi contenenti dati sensibili. Entra in gioco l'autenticazione biometrica, con la quale definisce il processo di verifica dell'identificazione di una persona sulla base di una sua caratteristica oppure di un suo tratto. Sostituire le password con una biometrica ha molti vantaggi, tra i quali non può essere dimenticata, nessuno può rubarla e non può essere trasferita ad un'altra persona. I tratti biometrici possono essere fisiologici oppure comportamentali. La biometria fisiologica, come il volto oppure le impronte digitali, è stata già introdotta con successo in molti dispositivi, come ad esempio FaceID nei nostri telefoni. La biometria comportamentale cattura un modello o un comportamento, come la firma o la verifica vocale. L'aspetto interessante della biometria comportamentale che essa può essere difficile da falsificare, ma allo stesso tempo anche difficile da modellare ed autenticare in modo robusto. Come accennato prima, esistono molti tipi di caratteristiche biometriche, ad esempio il volto, l'impronta digitale, l'iride, le labbra, il parlato, ecc. L'autenticazione è un processo a due fasi che prevede una fase di enrollment e poi l'autenticazione rispetto ai dati di registrazione. Questo processo è illustrato nella Fig. 1.4.

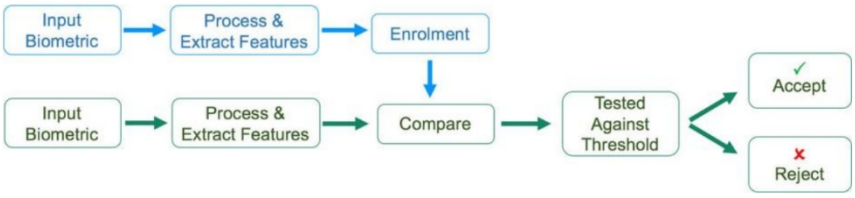

Figura 1.4: Panoramica dell'autenticazione biometrica, che prevede una fase di enrollment, in blu, e una fase di autenticazione, in verde.

Prendiamo in cosiderazione la lettura visiva delle labbra, dove quest'ultima è una tecnica per analizzare il parlato trasformando il movimento e/o l'aspetto concreto delle parti biometriche fisiche che compongono il discorso in caratteristiche visive. Recenti ricerche dimostrano che il labbro umano, da solo, è in grado di fornire informazioni suffiecienti sull'identità del suo propretario [1, 2]. L'aspetto (consistenza e forma) della regione labiale umana è unico e può essere considerato una caratteristica fisiologica. D'altra parte, il

movimento delle labbra durante l'enunciazione è influenzato dal modo in cui parla il parlante e ciò differisce da persona a persona, e, in più, può essere considerato una caratteristica comportamentale. Rispetto alle caratteristiche biometriche tradizionali, la caratteristica delle labbra prensenta i seguenti vantaggi. In primo luogo, la biometrica labiale è basata sul video, il che migliora la robustezza di un sistema di sicurezza garantendo "vivacità". In secondo luogo, il prezzo e l'utilità di una videocamera sono più accessibili rispetto ai costosi sensori utilizzati in alcuni sistemi biometrici. Infine, la biometrica labiale può essere facilmente integrata con altre caratteristiche biometriche come il volto, la voce, ecc. per costruire un sistema biometrico multimodale e fornire un livello di sicurezza molto elevato [3]. La sfida principale dell'uso della biometria labiale è come estrarre in modo efficacie le caratteristiche utili dal video che inquadra solamente la zona labiale. Lo schema di estrazione delle caratteristiche deve mantenere informazioni sufficienti per identificare il parlante e allo stesso tempo essere robusto alla variazioni causate dalla posa e dalla posizione del parlante rispetto alla telecamera, ecc. Normalmente, si utilizza tale approccio quando il suono non è disponibile oppure ci si ritrova in un ambiente rumoroso e quindi identificare una persona diventa abbastanza complesso se ci dovessimo basare sul suono. Questo metodo può essere applicato in varie applicazioni di sicurezza, ad esempio l'inserimento di password in sistemi di sorveglianza, per l'autenticazione di utenti, per il riconoscimento di parole e per persone con problemi di udito. L'autenticazione biometrica basata sul labbro (Lip-Base Biometric Authentication, LBBA) è il processo di autenticazione di una persona basato sui movimenti visivi delle labbra nel momento in cui si parla. Ad oggi, la LBBA non ha ottenuto un numero significativo di ricerche, soprattutto se paragonata ad altre biometrie come il volto, l'impronta digitale o la voce. Ci sono molti lavori in cui molti metodi proposti sono sporadiche e incoerenti, con risultati riportati su insiemi di dati piccoli o privati e metriche di risultato diverse che rendono difficile il confronto. L'autenticazione del relatore basato esclusivamente sulla funzione labiale (che viene denominata Visual Speaker Authentication, VSA) è stata studiata dagli anni '90. Approcci sofisticati che si basano sulle funzionalità labiali possono ottenere risultati di autenticazione affidabili (con un tasso di errore totale dimezzato inferiore all'1%) nello scenario a password fissa. Sebbene VSA a password fissa garantisca una doppia sicurezza, ovvero la password e l'aspetto del labbro/comportamento di conversazione, un attacco di riproduzione che utilizza un video pre-registrato può compromettere un tale sistema. Per resistere a tali attacchi di riproduzione e garantire la liveness, è stato proposto per la prima volta uno schema di password casuali [4].

1.7 Tecnologie utilizzate

In questa sezione verranno menzionate le tecnologie utilizzate nel corso del lavoro sul quale la presente trattazione è basata. È importante notare che se è stato utilizzato un unico linguaggio di programmazione, le librerie delle quali si è usufruito risultano essere varie ed eterogenee, in quanto ognuna di esse offre servizi differenti utili per portare a termine una specifica fase del progetto.

- **Linguaggi di programmazione**
  - **Python:** è l'unico linguaggio di programmazione utilizzato. E' stato scelto per la sua grande versatilità e per la sua vasta scelte di libreria e API che offre nell'ambito di Deep Learning.
- **Librerie e API**

  - **Tensorflow:** è la libreria software alla base dei modelli presenti in questa trattazione. È stata scelta per la sua grande duttilità e semplicità di utilizzo.
  - **Keras:** è l'API che è stata utilizzata per la realizzazione dei modelli neurali. Fa uso della libreria Tensorflow.
  - **OpenCV:** è la libreria utilizzata nell'ambito della computer vision. Per l'elaborazione delle immagini è la libreria più flessibile ed efficiente.
  - **NumPy:** è la libreria scelta per la manipolazione degli array multidimensionali. In questo progetto è ampiamente utilizzata insieme alla libreria OpenCV.

- **IDE e ambienti di sviluppo**

  - **Visual Studio Code**: è l'IDE utilizzato per le operazioni di image e video processing, per la preparazione dei dataset e implementazione del modello neurale.

Capitolo 2

Lavori correlati

In questo capitolo verranno citati alcuni lavori che hanno lo stesso input in comune, ovvero la zona labiale, ma affrontate con strategie diverse. Essendo l'obiettivo di questo lavoro quello di migliorare l'identificazione biometrica di un soggetto integrando la lingua parlata tramite la zona labiale escludendo la parte audio, questo capitolo verrà separato in due parti. La prima parte tratterà dei lavori che trattano la problematica dell'identificazione biometrica, e la seconda parte, il riconoscimento della lingua parlata.

2.1 Lavori correlati all'identificazione biometrica tramite zona labiale

Negli ultimi decenni, molti ricercatori hanno proposto vari metodi VSA (Visual Speaker Authentication), e questi lavori si possono approssimativamente suddividere in base al tipo di approccio che si utilizza, come: approccio basato su funzionalità *handcrafted-feature* e approccio basato sull'apprendimento automatico (*learning-based*). Per l'approccio basato su funzionalità handcrafted-feature, sono stati proposti vari tipi di rappresentazioni delle caratteristiche per descrivere le informazioni statiche (la forma e l'aspetto delle labbra) e dinamiche (il parlato) della biometria labiale. Le caratteristiche delle labbra più utilizzate per l'identificazione di una persona sono:

- Descrittori della forma labiale: descrittori di contorno geometrici [5, 6, 7] e parametri del modello di contorno ottenuti dall'Active Shape Model (ASM) [8];
- Descrittori di texture delle labbra: distribuzione dell'intensità della regione labiale [6] e visibilità di denti/lingua [5];
- Descrittori di movimento delle labbra: vettore di movimento del contorno labiale [9].

A breve, verranno descritti vari lavori di altri ricercatori riassumendo in modo veloce il loro metodo di approccio riguardo all'autenticazione tramite la zona labiale.

In questa sezione verranno menzionati lavori riguardo all'identificazione biometrica tramite la zona labiale per dare l'idea al lettore quali diversi approcci si possono utilizzare per il medesimo obiettivo. Si è pensato di dividere tali lavori in due fasce:

- *Lavori basati su modelli Hidden Markov*
- *Lavori basati su approccio Learning-based*

La maggior parte di questi lavori si basano principalmente sul fattore visivo escludendo l'informazione dell'audio del soggetto parlante.

2.1.1 Lavori basati su modelli Hidden Markov

In [10, 6, 7] hanno utilizzato i modelli Hidden Markov (HMM) per il riconoscimento del parlato e dell'oratore basato sulle labbra. HMM è il metodo di classificazione dei parlanti più utilizzato per le caratteristiche handcrafted. I lavori, sopra citati, utilizzando handcrafted-feature avevano dimostrato la possibilità dell'utilizzo della funzione labiale per l'autenticazione dell'oratore, tuttavia, le loro prestazioni non erano del tutto paragonabili a quelle dei recenti sofisticati approcci di autenticazione basati sul volto. Uno dei lavori che ha riscontrato maggior successo, riguardo all'utilizzo delle funzionalità handcrafted-feature, è stata quello pubblicato nel 2012 da Chan et al. [11]. La loro caratteristica era essenzialmente un descrittore di texture, che veniva chiamato Local Ordinal Constrast Pattern(LOCP), in cui tre piani ortogonali (TOP) sono stati impiegati nella loro funzione per considerare sia le informazioni spaziali che temporali. Un basso HTER (Half Total Error Rate) dello 0,36% è stato raggiunto sul database XM2VTS [12] con circa 300 parlanti, il che ha dimostrato l'alto potere discriminante della biometrica labiale.

Carlos M. Travieso et al. [13] lavorano su un nuovo ed efficace approcio di identificazione biometrica basato sulle labbra con il Discrete Hidden Model Kernel (DHMMK). Le labbra sono descritte da caratteristiche di forma (sia geometriche che sequenziali) su due diversi layout di griglia: rettangolare e polare. Queste caratteristiche sono poi modellate in modo specifico da un DHMMK e apprese da un modello SVM (Support-Vector Machine). Il loro approccio si divide in tre parti principali: l'estrazione delle labbra, la parametrizzazione DHMMK e la classificazione 2.1. La prima fase consiste nell'estrarre le labbra da un'immagine rilevando prima il volto, poi la bocca e infine le labbra. Il rilevamento del volto è ottenuto utilizzando il rilevatore polare di volti di Viola e Jones [14] che fornisce tassi di rilevamento del 99% sui tre set di dati utilizzati. Una volta rilevato il volto, il passo successivo è la localizzazione della bocca e l'eliminazione di eventuali effetti di contorno.

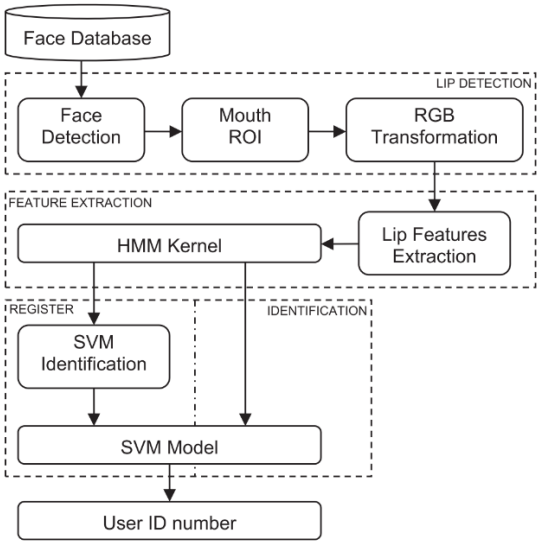

Figura 2.1: Schema a blocchi dell'approccio proposto da Carlos M. Travieso et al.

I loro esperimenti vengono condotti in una fase di cross-validation 10-fold su tre diversi set di dati, GPDS-ULPGC [15], Pie Face [16] e Rafd Face [17]. Il set di dati GPDS-ULPGC è stato raccolto da loro stessi ed è composto da molte persone estraendo per ognuna di esse dieci campioni, ottenendo in totale un dataset di 500 immagini. GPDS-ULPGC è organizzato in base al sesso delle persone, avendo il 54% di immagini di presenza maschile e per il 46% da quello femminile. Il campione, estratto da ogni soggetto, è un'immagine a colori di dimensioni 768 x 1024 pixel. Tale organizzazione del dataset ha dato come risultato il 99.8% di accuratezza. Il dataset Pie è scaricabile online [16] ed è composto da 68 soggetti, dove per ognuno di essi verranno estratte 11 campioni, ottenendo un totale di 748 immagini. Ogni campione è un'immagine a colori di dimensioni 200 x 300 pixel. La particolarità di aver utilizzato questo dataset era per via di immagini contenenti variazioni di illuminazione e diverse acconciature (ad esempio, barba o senza) dando come risultato il 97.13% di accuratezza. Infine il dataset Rafd è composto da un totale di 540 immagini con 60 soggetti differenti. La risoluzione di ogni immagine è 1024 x 681 pixel e il dataset offre 8 espressioni facciali per ogni soggetto. Utilizzo del dataset Rafd ha offerto, al lavoro di M. Travieso, il 98.10% di accuratezza al loro lavoro. Nel corso del loro lavoro non verranno proprio utilizzate le componenti audio ma solamente quelle visive.

2.1.2 Lavori basati con approccio Learning-based

Di recente, stanno diventando sempre più efficienti gli approcci di apprendimento automatico (*Intelligenza Artificiale*), come ad esempio la *codifica sparsa* e *le rete neurali profonde* dove hanno superato gli approcci precedentemente descritti in molte applicazioni di visione artificiale.

Il lavoro di Meng Liu et al. [18] introduce un benchmark AV-LB (*Audio-Video Lip biometrics*) basato sul Deep Learning, denominato DeepLip, realizzato con moduli convoluzionali unimodali video e audio e un modulo di fusione multimodale. I loro esperimenti dimostrano che DeepLip supera il sistema tradizionale di biometria labiale nella modellazione del contesto e ottiene miglioramenti relativi di oltre il 50% rispetto ai sistemi unimodali tramite la fusione audio-visiva, con tasso di errore pari allo 0,75% e all'1,11% rispettivamente sul set di dati di prova.

Anche nel lavoro di Feng Chenga et al. [4] si utilizza l'approccio learning-based, in particolare si soffermano su una problematica specifica, ossia si pongono l'obiettivo su come evitare attacchi di riproduzione. La loro idea è quella di proporre un nuovo schema di autenticazione visiva dei soggetti con un prompt di testo casuale, in breve permette l'autenticazione del soggetto tramite la zona labiale usando una password casuale (data dal sistema) che il soggetto dovrà pronunciare per essere autenticato. Per risolvere questo problema, è stata proposta una nuova rete neurale profonda composta da tre parti funzionali, vale a dire, il lip feature network, la rete d'identità e la rete di contenuti. Nella fase di *lip feature network*, una serie di unità residue 3D che possono rappresentare le caratteristiche statiche e dinamiche della biometria labiale in modo completo. Sono stati effettuati esperimenti per valutare le prestazioni della rete proposta sia nello scenario a password fissa che in quello con messaggi di richiesta casuali. I risultati dell'esperimento, è dimostrato che l'approccio proposto può ottenere prestazioni superiore dello scenario della password fissa confrontato con diversi approcci all'avanguardia. Inoltre, ottiene anche risultati di autenticazione soddisfacenti nello scenario dei testi di richiesta casuali, offrendo una accuratezza del 98.18%.

Invece, nel lavoro di Carrie Wright et al. [19] esplora l'idoneità dell'autenticazione basata sulle labbra come biometria comportamentale per i dispositivi mobili. Si è pensato a tale approccio perchè è particolamente adatta ai dispotivi mobili, il cui motivo è dovuto dal fatto che per catturare la zona labiale e effettuare un'autenticazione, basta utilizzare solamente una fotocamera frontale senza aver bisogno di un ulteriore hardware. Nel seguente lavoro, viene proposto il sistema, *LipAuth*, dove è stato esaminato in maniera rigoroso con i dati e le sfide del mondo reale che ci si può aspettare da una soluzione basata sulle labbra implementata su un dispositivo mobile, basandosi sull'approccio one-shot learning [20]. Successivamente, viene mostrato su come il sistema si comporta al di là di un protocollo chiuso, confrontando un nuovo protocollo aperto con un tasso di errore pari all'1.65% sul set di dati XM2VTS offrendo un'accuratezza del 95.41%. All'interno di esso, vengono raccolti nuovi set di dati, qFace e FAVLIPS, che fanno progredire il campo consentendo di testare sistematicamente il contenuto e le quantità di dati necessari per l'autenticazione biometrica basata sulle labbra e di evidenziare le aree problematiche per il lavoro futuro. Il set di dati FAVLIPS è stato progettato per simulare alcune delle sfide più difficili che ci si protrebbe aspettare in uno scenario di implementazione e comprende contenuti aventi problematiche, come la mimica e problemi di illuminazione. L'approccio di Krzysztof Wrobel et al. [21] consente di utilizzare le caratteristiche composte - i cosiddetti coefficienti *Sim* - e i punti di riferimento che determinano l'area in cui le caratteristiche biometriche devono essere ricercate. Le caratteristiche biometriche composte sono associate ad appropriati coefficienti di somiglianza. Questo approccio porta vantaggi significativi: il livello di riconoscimento degli oggetti è più alto rispetto al metodo

basato sui dati grezzi. In questo lavoro, viene proposto un nuovo ed efficace approccio di riconoscimento biometrico basato sulle labbra con la rete neurale probabilistica (PNN). Nella prima fase, l'area delle labbra viene ristretta a una regione di interesse (ROI) e nella seconda fase le caratteristiche estratte dalla ROI vengono modellate in modo specifico da algoritmi di elaborazione delle immagini dedicati, come in questo caso HOG e Haar-like, utilizzati successivamente come input della rete neurale. Nella figura 2.2 viene descritto il loro approccio in tre fasi: Lip detection, Lip Feature extraction e Classification Method.

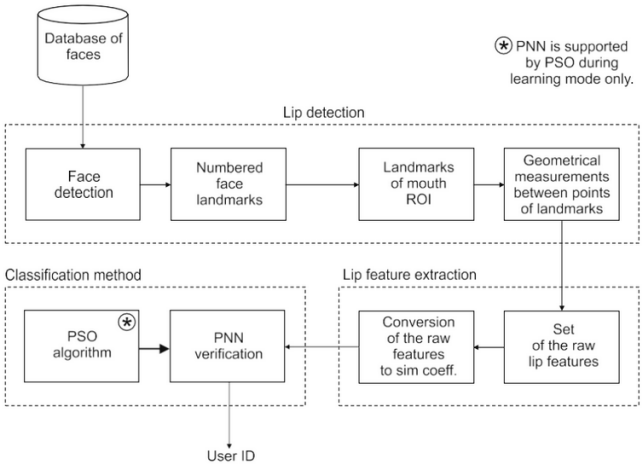

Figura 2.2: Schema a blocchi dell'approccio proposto da Krzysztof Wrobel et al.

Tutti gli esperimenti sono stati confermati con cross-validation 10-fold su tre set di dati: Multi-PIE Face, PUT e un dataset da loro creato dando come accuratezza media di classificazione dell'86.95%, dell'87.14% e dell'87.26% rispettivamente. Tali risultati sono stati ottenuti grazie alla tecnica Particle Swarm Optimization (PSO). In letteratura, molti lavori hanno utilizzato immagini labiali bidimensionali (2D) per riconoscere l'oratore in un contesto dipendente dal testo.

Tuttavia, il labbro 2D soffre facilmente di vari orientamenti del viso. Con il lavoro di Jianrong Wang et al. [22] presentano una nuova rete di movimento labiale 3D end-to-end (3LMNet) utilizzando il movimento labiale 3D a livello di frase (S3DLM) per riconoscere i parlanti sia nel contesto indipendente dal testo che in quello dipendente dal testo. Viene proposto un nuovo modulo di feedback regionale (RFM) per ottenere attenzioni in diverse regioni del labbro. Loro presentano due metodi, vale a dire, coordinare la trasformazione e la correzione della postura del viso per effettuare il preprocessing del set di dati LSD-AV, che contiene 68 parlanti e 146 frasi per parlante. I risultati del loro lavoro sul set di dati LSD-AV dimostrano che 3LMNet è superiore ai modelli di riferimento, ovvero LSTM, VGG-16 e ResNet-34, e supera lo stato dell'arte, con un'accuratezza del 99.1%, utilizzando l'immagine labiale 2D e la faccia 3D.

Infine, con Chen-Zhao Yang et al. [23], si affronta la tematica del DeepFake. Tale tecnica, se usata per scopi loschi, può compromettere seriamente i tradizionali sistemi di autenticazione basati sul viso o sulle labbra. Per difendersi da sofisticati attacchi DeepFake, in questo lavoro viene proposto un nuovo schema di autenticazione dei soggetti basato sulla rete neurale convoluzionale profonda (DCNN). La rete proposta è composta da due parti funzionali, vale a dire, la rete di estrazione delle caratteristiche fondamentali (FFE-Net) e la rete di estrazione e classificazione delle caratteristiche del labbro rappresentativo (RC-Net). Nella Figura 2.3 possiamo notare l'architettura del sistema di autenticazione proposta.

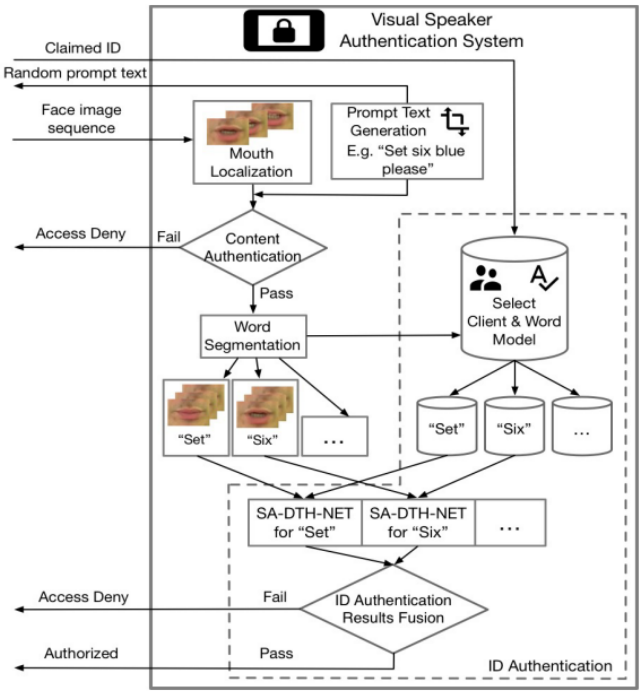

Figura 2.3: Architettura dell'approccio proposto da Cheng-Zhao Yang et al.

La FFE-Net fornisce le informazioni fondamentali per l'autenticazione dei soggetti. Poichè la forma statica delle labbra e l'aspetto delle labbra sono vulnerabili agli attacchi DeepFake, il movimento dinamico delle labbra è enfatizzato nella rete FFE. RC-Net estrae caratteristiche labiali di alto livello che discriminano gli impostori umani mentre catturano lo stile di conversazione del proprietario. Viene progettato uno schema di apprendimento multi-task e la rete proposta viene formata end-to-end. Gli esperimenti vengono effettuati sui set di dati GRID e MOBIO dove hanno dimostrato che l'approccio proposto è in grado di ottenere un risultato di autenticazione accurato contro impostori umani ed è molto più robusto contro gli attacchi DeepFake rispetto ad alcuni algoritmi di autenticazione degli interlocutori visivi all'avanguardia.

2.2 Lavori correlati al riconoscimento della lingua parlata

Uno dei primi lavori che ha affrontato la problematica del riconoscimento della lingua parlata tramite analisi visiva in modo diretto, è stato quello di Jacob L. Newman et al. [24] pubblicato nel 2010. Il loro lavoro si è concentrato sul riconoscimento di due specifiche lingue, inglese e francese, elaborando dei modelli basati su *Support Vector Machine (SVM)* o macchine a vettori di supporto e sugli *Hidden Markov Model (HMM)* o modelli di Markov nascosti. Hanno utilizzato il dataset *VLID* che consisteva in registrazioni di 21 soggetti. Ognuno di loro conosceva due o più lingue, tra le quali una era la lingua madre del soggetto e le altre erano lingue dove egli aveva discrete conoscenze. Ogni soggetto doveva leggere un copione scelto davanti ad una telecamera, evitando di coprire il viso, in tutte le lingue di cui era a conoscenza. L'obiettivo del lavoro era tentare di riconoscere la lingua francese e quella inglese. Nonostante il loro lavoro abbia prodotto un risultato basso, ossia il 34% di accuratezza, essi dimostrarono che l'automatizzazione del riconoscimento della lingua parlata è possibile, cercando di aumentare l'ampiezza del dataset e aggiungendo altre lingue. Infatti,

successivamente fecero un ulteriore lavoro nel 2012 [25] in cui aumentarono il numero di video e aggiunsero altre 2 lingue, ossia l'arabo e il tedesco, presi dai dataset *United Nations 1* e *United Nations 2* dove entrambi contenevano tracce audio e video dei soggetti. L'organizzazione del dataset è uguale a quello citato precedentemente, ovvero per ogni soggetto aveva a disposizione un copione scelto dove esso doveva leggere in tutte le lingue di cui aveva conoscenze.

Nel corso del loro lavoro furono realizzati esperimenti sia in contesto *speaker-dependent* sia in *speaker-independent*. Nel primo caso il modello è in grado di classificare la lingua in cui un soggetto multilingua sta parlando, mentre, nel secondo caso, i soggetti sono tutti diversi e ciascun soggetto parla una sola lingua. Per la valutazione del modello è stato utilizzato, per la divisione del dataset, la *cross-validation 19-fold* dove per ognuno dei 19 soggetti veniva tenuto fuori dal set di training e usato per il testing del modello a turno. Per ogni soggetto utilizzato nel set di testing, veniva diviso in sequenze da 1, 3, 7, 30, e 60 secondi. Difatti il lavoro non ebbe un notevole miglioramento tenendo in considerazione quello precedente ma ci fu un miglioramento riguardo al tasso di errore, nei modelli implementati, sulla lingua araba e inglese su 19 soggetti e usando registrazioni del parlato di 30 secondi. Infine, un lavoro di recente pubblicazione, è quello di Triantafyllos Afouras et al. rilasciato nel 2020 [26]. L'obiettivo del loro lavoro è l'implementazione di modelli neurali in grado di identificare una lingua specifica semplicemente facendo riferimento ai movimenti delle labbra di un soggetto. In questo studio hanno fatto riferimento a due dataset: *LRS3-Lang+* e *VoxCeleb2*, di cui il primo è composto da video brevi (dalla durata di 6 secondi ognuno) multilingua raccolti dalle conferenze TEDx che comprende parlanti di tredici diverse lingue: il giapponese, il coreano, il greco, il polacco, il russo, il turco, il cinese, l'italiano, l'arabo, il francese, l'inglese, lo spagnolo e il portoghese. Il secondo non contiene etichette linguistiche, ma l'identità dei parlanti e la loro nazionalità sono note. I video contenuti nel dataset *VoxCeleb2* sono stati presi dalla piattaforma di YouTube quindi entrambi i dataset presi in esame, sono stati raccolti online quindi, in confronto ai lavori precedentemente descritti, i soggetti contenuti non hanno vincoli sul parlato. Gli esperimenti sono stati eseguiti utilizzando diversi strategie iniziando con le *Time-Delay Neural Networks* (TDNN), per continuare con approcci misti che comprendessero anche l'utilizzo delle *Bidirection Long Short-Term Memory* (BLSTM), per poi terminare con delle reti caratterizzate da tre strati di BLSTM. L'addestramento dei modelli è stato svolto principalmente sul primo dataset, invece la fase di testing sono stati utilizzati sia il primo che il secondo. Tale approccio ha permesso di non sovrapporre soggetti che venivano utilizzati sia in fase di traning che di testing producendo un risultato affidabile alla realtà. I modelli implementati, in fase di addestramento, prenderanno in input sequenze di video facciali in cui sono stati precedentemente modellate, tramite la rete neurale *ResNet18*, il movimento della labbra. L'esperimento riguardo allo strato bidirezionale LSTM ha raggiunto come risultato 84% di accuratezza, eseguendo i test sul dataset LRS3-Lang+ su sequenze di 30 secondi, e con il dataset VoxCeleb2 si è ottenuto il 67.4% con sequenze di video da 10 secondi.

2.3 Approccio utilizzato

Lo scopo di questo lavoro è cercare di ottenere un incremento nei risultati dati dal modello che identifica un soggetto integrando i risultati ottenuti riguardo al riconoscimento della lingua parlata del medesimo soggetto, quindi si parla di una fusione tra queste due problematiche citate precedentemente. Lo sviluppo del seguente lavoro, in confronto a quelli citati sopra, presenta alcune differenze riguardo agli approcci utilizzati, che verranno definite in seguito nei successivi capitoli. Tali differenze sono le seguenti:

- **Dataset proprietario:** il dataset utilizzato è unico, nel senso sono stati selezionati i video per lo scopo di questo lavoro, e a differenza del dataset citati poco prima, all'interno verranno considerate le seguenti lingue per la classificazione: italiano, inglese, tedesco, spagnolo, olandese, russo, giapponese e francese. La selezione dei video è stata elaborata seguendo determinati criteri, che verranno descritti successivamente nel Capitolo 3.

- **Deep Learning:** a differenza degli studi citati, nel seguente lavoro sono stati utilizzati modelli di Deep Learning. Tale approccio comporta all'utilizzo diretto delle sequenze video e, in questo specifico lavoro, non verranno trattate le componenti audio. I video verranno convertiti in dati binari in modo tale da renderli comprensibili al modello scelto, dove quest'ultimo verrà utilizzato per entrambe le problematiche.

Capitolo 3

Metodo proposto

Nel seguente capitolo saranno descritte le fasi di lavoro attraversate, partendo dalla raccolta video per l'organizzazione del dataset e della motivazione per cui non si è utilizzato uno già esistente, il processo di estrazione dei sottovideo e della zona labiale e infine la fase di preprocessing realizzata con lo scopo di rendere i dati accessibili per il modello neurale realizzato e infine arrivando all'implementazione di quest'ultimo. Prima saranno descritti, in modo teorico, alcuni concetti e metodi utilizzati in modo tale da dare un'idea al lettore di cosa si sta trattando.

3.1 Nozioni Teoriche

3.1.1 Rete Neurale

Le reti neurali artificali (ANN), più comunemente chiamate reti neurali (NN), sono sistemi di calcolo ispirati alle reti neurali biologiche che costituiscono il cervello umano.

Breve storia

Le reti neurali possono sembrare un argomento nuovo ed entusiasmante, ma il campo stesso non è affatto nuovo. Si pensa che nel 1943 Warren McCulloch e Walter Pitts crearono il primo modello di neurone artificiale, basato su algoritmi e sulla logica del *threshold*, ovvero la logica che permette di convertire un input di natura continua in un output di natura discreta. Questo modello, ancora alla base delle più recenti versioni di reti neurali, era basato su una funzione di attivazione, versione matematica del potenziale di azione che permetteva al neurone di assumere i due valori di attivo e silente, stati determinati dagli input ricevuti dal neurone e smorzati dai pesi sinaptici che legano il neurone artificiale alle unità neurali degli strati precedenti. Questa modellizzazione fu essenziale per i successivi sviluppi nell'ambito delle reti neurali, che permisero nel 1958 l'implementazione da parte di Rosenblatt del percettrone a singolo strato, capace di eseguire operazioni di classificazione binaria. L'opera di Rosenblatt stimola una quantità di studi e ricerche che dura per un decennio, e suscita un vivo interesse e notevoli aspettative nella comunità scientifica, destinate tutta via ad essere notevolmente ridimensionate allorchè nel 1969 Marvin Minsky e Seymour A. Papert, nell'opera *Perceptrons. An Introduction to Computational Geometry*, mostrano i limiti operativi delle semplici reti a due strati basate sul percettrone, e dimostrano l'impossibilità di risolvere per questa via molte classi di problemi, ossia tutti quelli non caratterizzati da separabilità lineare delle soluzione: questo tipo di rete neurale non è ab-

bastanza potente; non è infatti neanche in grado di calcoare la funzione *or esclusivo* (XOR). A causa di queste limitazioni, al periodo di euforia dovuto ai primi risultati della cibernetica (come veniva chiamata negli anni '60) segue un periodo di diffidenza durante il quale tutte le ricerche in questo campo non ricevono più alcun finanziamento dal governo degli Stati Uniti d'America; le ricerche sulle reti tendono, di fatto, a ristagnare per oltre un decennio, e l'entusiasmo iniziale risulta fortemente ridimensionato. Fino ad arrivare negli anni '80, dove fu dimostrato che le reti neurali potevano avere un ruolo fondamentale nella risoluzione di problemi contemporanei. Per questo si iniziò nel 1989 con la creazione della prima rete neurale convoluzionale. Le operazioni di convoluzione permeettono di estrarre caratteristiche salienti dalle immagini, non alterando la struttura bidimensionale delle stesse. Esse sono svolte attraverso un filtro immaginabile come una matrice i cui valori sono moltiplicati per i valori che costituiscono i pixel dell'immagine. Proprio attraverso queste moltiplicazioni si ottiene l'estrazione delle caratteristiche. L'introduzione di questo nuova tipologia di rete permise per la prima volta a un algoritmo, di classificare in dieci classi delle immagini che ritraevano delle cifre scritte a mano.

Definizione

Una *rete neurale artificiale* (ANN "Artificial Neural Network" in inglese), normalmente chiamata solo "rete neurale" (NN "Neural Network" in inglese), è un modello matematico/informatico di calcolo basato sulle reti neurali biologiche. Tale modello è formato da un gruppo di interconnessioni di informazioni costituite da neuroni artificiali e processi che utilizzano un approccio di connessionismo di calcolo 3.1. Nella maggior parte dei casi una rete neurale è un sistema adattivo che cambia la proprio struttura in base a informazioni esterne o interne che scorrono attraverso la rete stessa durante la fase di apprendimento. In termini pratici le reti neurali sono strutture non-lineari di dati statistici organizzate come strumenti di modellazione. Esse possono essere utilizzate per simulare relazioni complesse tra ingressi e uscite che altre funzioni analitiche non riescono a rappresentare. Una rete neurale riceve segnali esterni su uno strato di nodi (unità di elaborazione) di ingresso, ciascuno dei quali è collegato con numerosi nodi interni, organizzati in più livelli. Ogni nodo elabora i segnali ricevuti e trasmette il risultato a nodi successivi.

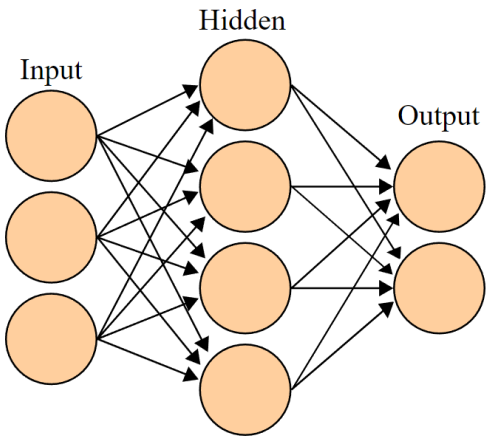

Figura 3.1: Una rete neurale artificiale è un'interconnessione di un gruppo di nodi chiamati neuroni

3.1.2 RNN (Recurrent Neural Network)

Gli esseri umani non iniziano il loro pensiero dal nulla. Ad esempio, mentre il lettore sta leggendo questo lavoro egli capisce ogni parola in base alla sua comprensione delle parole precedenti, quindi i pensieri hanno una persistenza. Le reti neurali tradizionali non possono farlo, e sembra una grave problematica. Ad esempio, si immagina di voler classificare il tipo di evento che sta accadendo in ogni punto di un film. Una rete neurale tradizionale non riesce a tener in considerazione gli eventi precedenti per informare quelli successivi. Le *reti neurali ricorrenti* risolvono questo problema. Sono reti 3.2 con loop al loro interno, che consentono alle informazioni di persistere.

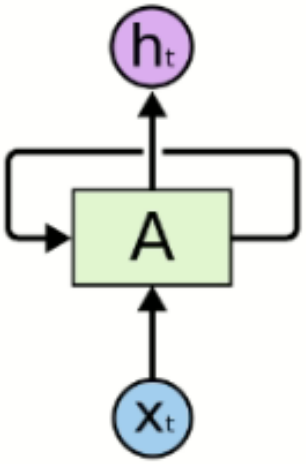

Figura 3.2: Struttura rete neurale [27]

Nella figura sopra, un pezzo di rete neurale A, esamina input x_t e restituisce un valore h_t. Un ciclo consente di mantenere l'informazione in vita. Le reti ricorrenti prevedono, in sostanza, collegamenti all'indietro o verso lo stesso livello. Questa caratteristica rende questo tipo di rete neurale molto interessante, perchè il concetto di ricorrenza introduce intrinsecamente il concetto di memoria di una rete. In una rete RNN, infatti, l'output di un neurone può influenzare se stesso, in uno step temporale successivo o può influenzare neuroni della catena precedente che a loro volta interferiranno con il comportamento del neurone su cui si chiude il loop. Ovviamente non esiste un solo modo di implementare una rete RNN, infatti nel tempo sono stati proposti e studiati diversi tipi di rete RNN tra le più note ricordiamo quelle che si basano sulle LSTM (Long Short Term Memory) e le GRU (Gated Recurrent Units).

3.1.3 Long Short Term Memory (LSTM)

LSTM o *Long Short Term Memory*, è una rete neurale artificale utilizza nei campi dell'Intelligenza Artificiale e del Deep Learning. A differenza delle reti neurali *feedforward* standard, LSTM ha connessioni di feedback. Le reti LSTM sono un tipo speciale di RNN, in grado di apprendere dipendenze a lungo termine e sono estremamente utili nella risoluzione di enormi problemi. Sono stati introdotti da Hochreiter e Schmidhuber nel 1997 e sono stati raffinati e resi popolari da molti ricercatori.

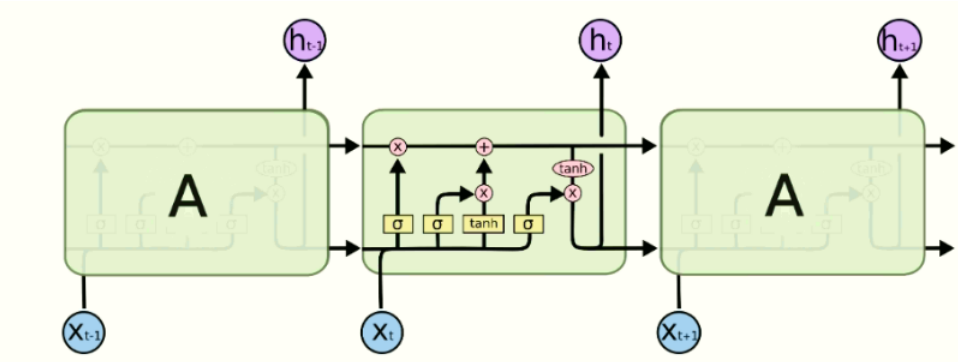

Figura 3.3: Modulo ripetitivo di una rete LSTM [27]

Le celle neurali di questo tipo di rete incorporano dei *gate* o porte interne che regolano il flusso di informazioni. Essi consistono unicamente in operazioni matematiche eseguite sui dati che fluiscono attraverso la cella di calcolo. Il loro scopo ultimo è quello di trattenere le informazioni ritenute utili e di scartare quelle giudicate superflue. L'unica differenza che separa le reti RNN da quelle LSTM risiede nei gate e nelle operazione da esse associate, dove quest'ultimi permettono alla rete di avere memoria, che si concretizza nel *cell state*, che custodisce la memoria a lungo termine e nell'*hidden state* correlata a una memoria a breve termine. Il passaggio dei dati attraverso i gate avviene attraverso alla funzione di attivazione sigmoidea che permette di attribuire ai dati un grado di importanza, dove se il dato passante attraverso al gate assumerà il valore 0 allora sarà giudicato come dato da scartare, invece se qualsiasi dato fosse associato il valore 1 sarebbe catalogato come molto importante. I gate caratterizzanti le celle, in ordine di attraversamento sono: il *forget gate*, l'*input gate* e l'*output gate*.

Forget gate

Il *forget gate*, il cui funzionamento è mostrato in Figura 3.4, decide quale informazione deve essere trattenuta e quale deve essere scartata. Per fare ciò, l'informazione rappresentata dall'output della cella precedente, che consiste nell'hidden state, viene è unita all'input della cella corrente e il risultato di questa combinazione è trasportato attraverso la funzione sigmoidea. Più i valori in output sono vicini ai due estremi 0 e 1, tanto più quell'informazione sarà ritenuta rispettivamente irrilevante o importante. L'equazione che descrive il forget gate è:

$$f_t = \sigma(W_f * [h_t - 1, x_t] + b_f) \tag{3.1}$$

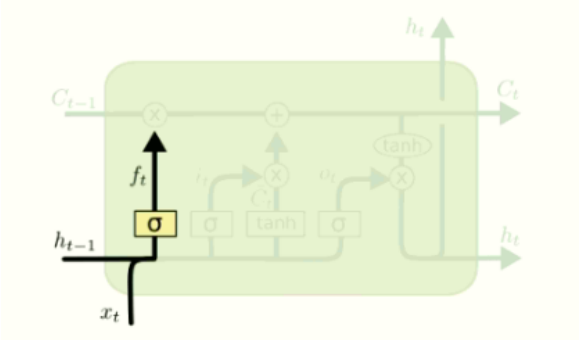

Figura 3.4: Funzionamento del forget gate [27]

Input gate

Il funzionamento dell'*input gate* è mostrato nella Figura 3.5. Esso consente di aggiornare lo stato della cella, ovvero, la condizione della memoria a lungo termine della cella. L'aggiornamento avviene in due fasi:

1. L'output della cella precedente e l'input della cella corrente sono combinati e la loro combinazione è data in input alla funzione sigmoidea, decidendo di conseguenza quali valori saranno aggiornati. Successivamente, uno strato *tanh* crea un vettore di nuovi valori candidati, C_t, che potrebbe essere aggiunto allo stato.
2. A questo punto si moltiplicano i risultati delle due funzioni, in modo tale che l'output della funzione sigmoidea decida quale informazione dell'output della funzione tangente sarà giusto trattenere.

Le formule matematiche che descrivono il funzionamento dell'*input gate* sono le seguenti:

$$i_t = \sigma(W_i * [h_t - 1, x_t] + b_i) \tag{3.2}$$

$$C_t = \tanh(W_C * [h_t - 1, x_t] + b_C) \tag{3.3}$$

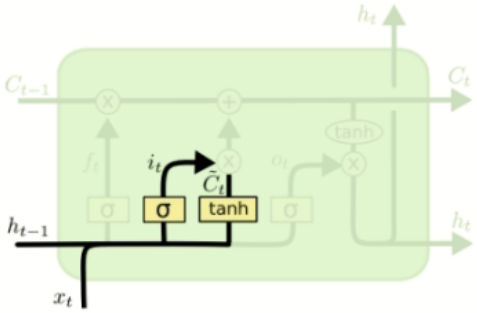

Figura 3.5: Funzionamento dell'input gate [27]

Output gate

Il terzo e ultimo gate attraversato dalle informazioni attraverso la cella di calcolo consiste nell'*output gate*, il cui funzionamento è mostrato in Figura 3.6 che decide quale sarà il prossimo *hidden state*. Per calcolarlo

lo stato nascosto della cella precedente e l'input corrente divengono l'input di una funzione sigmoidea, poi il *cell state* calcolato nella fase appena precedente diviene l'input per una funzione tangente, e alla fine gli output delle due funzioni sono moltiplicati generando il nuovo *hidden state*. La formula che descrive il suo funzionamento è:

$$o_t = \sigma(w_o * [h_t - 1, x_t] + b_o) \tag{3.4}$$

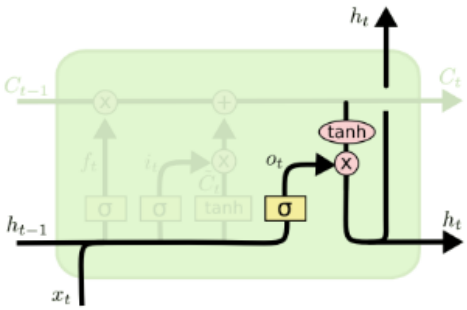

Figura 3.6: Funzionamento dell'output gate [27]

Tuttavia, esistono diverse variazioni della LSTM originaria, che puntano ad integrare un numero maggiore o tutti gli stai nascosti calcolati nel corso dell'iterazione, per effettuare la predizione finale del modello. Per rendere più facile la comprensione delle formule, citate poco prima, verranno descritte le notazioni usate nelle equazioni che descrivono la funzione dei gate:

$$i_t : rappresenta\ l'input\ gate \tag{3.5}$$

$$f_t : rappresenta\ il\ forget\ gate \tag{3.6}$$

$$o_t : rappresenta\ l'\ output\ gate \tag{3.7}$$

$$\sigma : rappresenta\ la\ funzione\ sigmoidea \tag{3.8}$$

$$w_x : parametri\ per\ eseguire\ le\ operazioni\ al\ gate\ x\ dove\ x\ assume\ i\ valori\ f,\ i,\ oppure\ o\ a\ seconda\ che\ ci\ si\ trovi\ al\ forget,\ input\ o\ output\ gate. \tag{3.9}$$

$$h_t - 1 : rappresenta\ l'\ output\ del\ precedente\ blocco\ lstm \tag{3.10}$$

$$x_t : rappresenta\ l'\ input\ del\ blocco\ lstm\ corrente \tag{3.11}$$

$$b_x : parametri\ aggiuntivi,\ variabili\ a\ seconda\ del\ gate. \tag{3.12}$$

3.1.4 Reti neurali ricorrenti in ambito visivo

Le reti neurali ricorrenti possono essere utilizzate anche nel caso in cui i dati raccolti siano di natura visiva, quindi nei casi in cui si debbano realizzare operazioni di classificazioni in relazione a video, sequenze di immagini oppure operazioni di riconoscimento della scrittura. Il motivo dell'introduzione delle reti neurali ricorrenti in contesti visivi, porta la necessità di operare delle modifiche sulla struttura originaria della rete. Tale necessità porta alla creazione delle *Convolutional Recurrent Neural Network (CRNN)* o reti neurali ricorrenti convoluzionali, che, oltre alle operazioni tipiche delle reti neurali classiche prevedono le operazioni di convoluzione, come illustrato in figura 3.7.

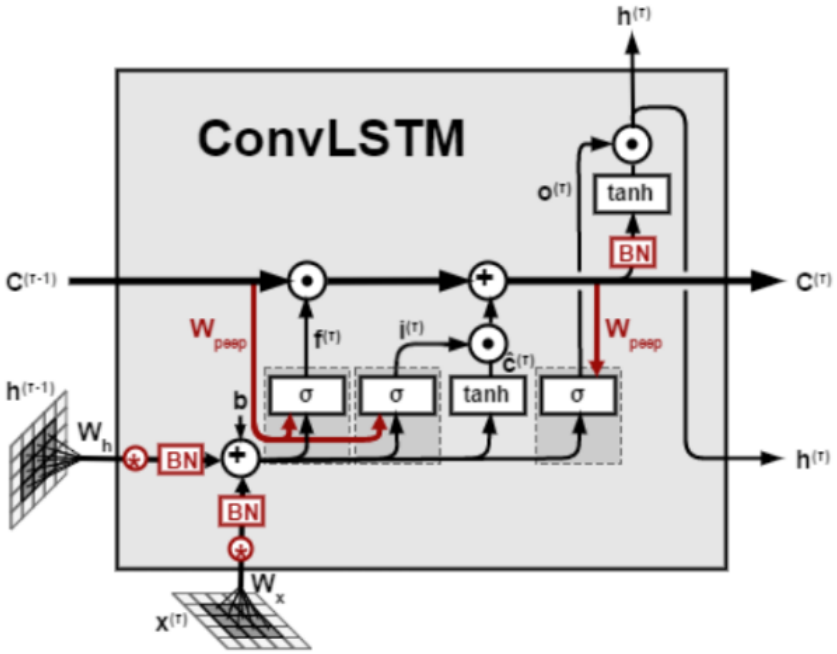

Figura 3.7: Architettura di una LSTM convoluzionale

Attraverso questa fusione diviene possibile sfruttare le potenzialità offerte da entrambi gli approcci:

- **Convoluzione:** le operazioni di convoluzione permettono di avere le caratteristiche principali delle immagini che compongono le sequenze sulle quali la rete sta effettuando l'addestramento
- **Ricorrenza:** la ricorrenza permette, attraverso i gate delle celle, di aggiungere sequenzialitàm, e quindi di permettere la contestualizzazione e la fusione delle caratteristiche individuate grazie alle operazioni di convoluzione.

ConvLSTM è un tipo di rete neurale ricorrente per la previsione spazio-temporale che presenta strutture convoluzionali nelle transizioni input-to-state e state-to-state. Il ConvLSTM determina lo stato futuro di una determinata cella nella griglia dagli input e dagli stati passati dei suoi vicini locali. Ciò può essere facilmente ottenuto utilizzando un operatore di *convoluzione* nelle transizioni da stato a stato e da input a stato 3.8.

La *kernel size* indica la dimensione l'altezza e la larghezza della finestra di convoluzione in formato 2D. Un ConvLSTM con un kernel grande dovrebbe essere in grado di catturare movimenti più veloci mentre

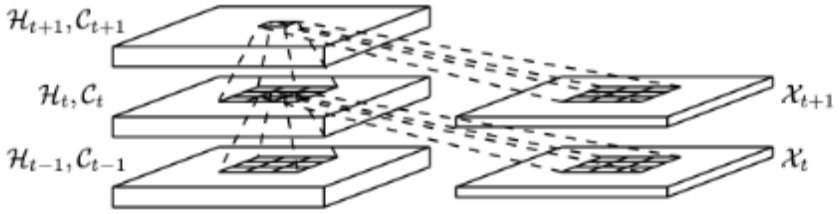

Figura 3.8: Struttura interna del ConvLSTM

con uno più basso può catturare movimento lenti. Per garantire che gli stati abbiano lo stesso numero di righe e lo stesso numero di colonne degli input, è necessario il riempimento prima di applicare l'operazione di convoluzione.

3.2 Creazione dataset

L'organizzazione del dataset è stato elaborato unicamente per questo lavoro. La motivazione di tale scelta ha permesso di organizzare il dataset in base alle strategie (*lingua, sesso, età*) che sono stati imposti e che verranno approfondite successivamente. Un'altra motivazione è che molti dei dataset già esistenti non erano reperibili gratuitamente online o del tutto e quei pochi disponibili non erano organizzati in modo efficiente per lo scopo prefissato. Andando nello specifico, i dataset citati nel Capitolo 2 nei *"Lavori correlati all'identificazione biometrica tramite zona labiale"*, ovvero *Pie*, *XM2VTS*, *GPDS-ULPGC*, non sono reperibili online e non facevano alcun riferimento alla lingua parlata dei soggetti presenti. In più, venivano raccolti solamente delle immagini per il loro scopo, quindi ben diverso da questo lavoro perchè si utilizzano i dati in formato video. Stessa identica ragione per i dataset citati nei *"Lavori correlati al riconoscimento della lingua"*. La maggior parte di essi non erano disponibili online ed erano composti da pochi soggetti diversi. Per queste ragioni si è deciso di elaborare un *"dataset proprietario"* per avere un numero bilanciato di soggetti organizzati per lingua, una discreta qualità per tutti i video e che rispettino dei criteri prefissati per questo lavoro, e infine che in futuro possa essere migliorato per altri lavori che hanno il medesimo obiettivo.

3.2.1 Raccolta dati

I dati, che saranno sottoposti a questo lavoro, sono semplicemente dei video dove è presente una persona parlante. Per questo lavoro, sono stati scelti dei video presenti su varie piattaforme online (come ad esempio YouTube) rispettando alcuni requisiti per evitare meno problemi e fornire un'accuratezza elevata con le strategie che verranno introdotte successivamente. Tali requisiti usati per la selezione dei video sono i seguenti:

- La durata del video deve essere sufficientemente lunga, o almeno di una durata superiore di 5 minuti;
- Il video deve catturare per la maggior parte della sua durata un'unica persona parlante all'interno di esso;
- Il soggetto all'interno del video si deve posizionare frontalmente davanti alla telecamera, in modo tale da rendere il suo viso pienamente visibile e con poco discostamento durante il parlato;
- La qualità del video deve essere preferibilmente con un discreto numero di framerate (dai 30 ai 60 fps) e un'alta risoluzione (vicino ad 720p).

I video ottenuti sono stati suddivisi per ciascuna lingua in base a due parametri, ossia età del soggetto e sesso, e cercando di avere una quantità equilibrata di video per ciascuna lingua. Per identificare i video è stata utilizzata una precisa nomenclatura (*lingua_sesso_età_id-unico_framerate*), che permette di individuare le caratteristiche di ciascun video:

- **Lingua:** è un numero che va da 1 a 8.
    1. Italiano
    2. Inglese
    3. Tedesco
    4. Spagnolo
    5. Olandese
    6. Russo
    7. Giapponese
    8. Francese
- **Sesso:**
    1. Uomo
    2. Donna
- **Età:**
    1. Meno di 30 anni
    2. Più di 30 anni
- **Etichetta del video:** numero per indicare univocamente il video nella sua categoria
- **Framerate:** numero che rappresenta il framerate del filmato

Come descritto prima, i video sono stati scaricati principalmente su YouTube da persone diverse senza un processo standardizzato, la loro qualità e la loro durata sono molto variabili. La maggior parte dei video contenuti nel dataset hanno come risoluzione *720p*, ma ci saranno alcuni con qualità inferiore e superiore, e in più gli script per l'estrazione dei sottovideo e della zona labiale, che verranno descritti successivamente, non andranno a modificare la qualità di essi. Di conseguenza tale dataset raccolto, in futuro, può essere reso molto più ampio e/o migliorato ulteriormente.

In totale i video raccolti sono 256 dove ognuno di essi si ha una persona diversa, tali video sono divisi in 32 per ogni lingua ottenendo un dataset bilanciato 3.2.1.

Figura 3.9: Video di lingua italiana con un framerate uguale a 30, con soggetto di sesso maschile e di età inferiore ai trent'anni

Totali Video Interi					
	Uomo		**Donna**		
	Under 30	**Over 30**	**Under 30**	**Over 30**	**Totale**
Italiano	8	6	10	8	**32**
Inglese	7	6	7	12	**32**
Tedesco	8	6	9	9	**32**
Spagnolo	7	11	6	8	**32**
Olandese	6	17	5	4	**32**
Russo	2	14	4	12	**32**
Giapponese	9	8	6	9	**32**
Francese	8	9	8	7	**32**
					256

Tabella 3.1: Tabella riassuntiva dei video presi in considerazione

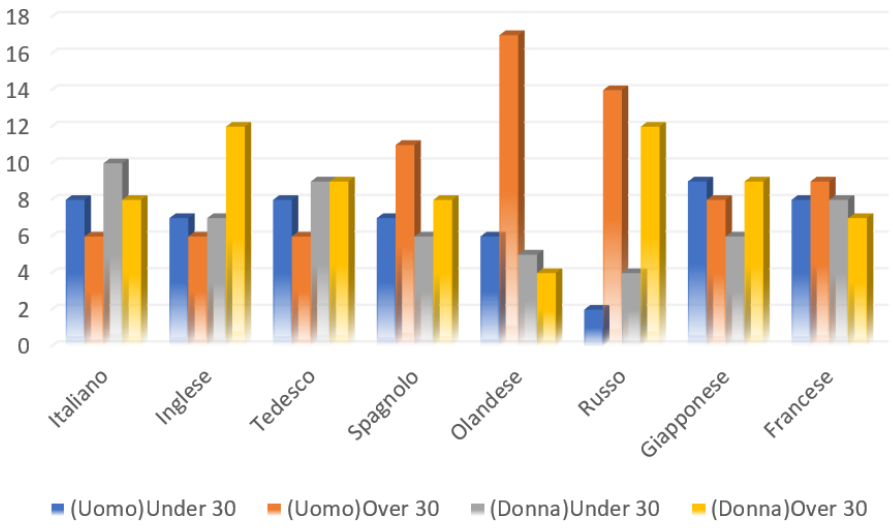

Figura 3.10: Organizzazione del dataset dei video interi

3.2.2 Estrazione dei sottovideo e della zona labiale

Avendo a disposizione il dataset organizzato e bilanciato come descritto prima, si prosegue alla fase di estrazione dei *sottovideo*. Durante questa fase, a ciascun video verranno estratte 5 sottovideo dalla durata di 10 secondi presi in tempi diversi. La scelta di tale strategia è dovuta dal fatto che si voleva avere una giusta quantità di materiale dalla stessa durata in modo tale da ottenere un risultato coerente. L'algoritmo in questione viene denominato *"extract_video.py"* scritto totalmente in linguaggio Python e concettualmente sfrutta la distanza euclidea per stabilire la distanza tra il 62-esimo e il 67-esimo landmark per il rilevamento della faccia del soggetto per ogni video da tagliare, successivamente viene eliminata completamente la traccia audio per ogni sottovideo essendo che si ha solamente bisogno del immagine visiva e infine tramite la libreria ffmpeg i sottovideo vengono salvati in formato *.mp4*.

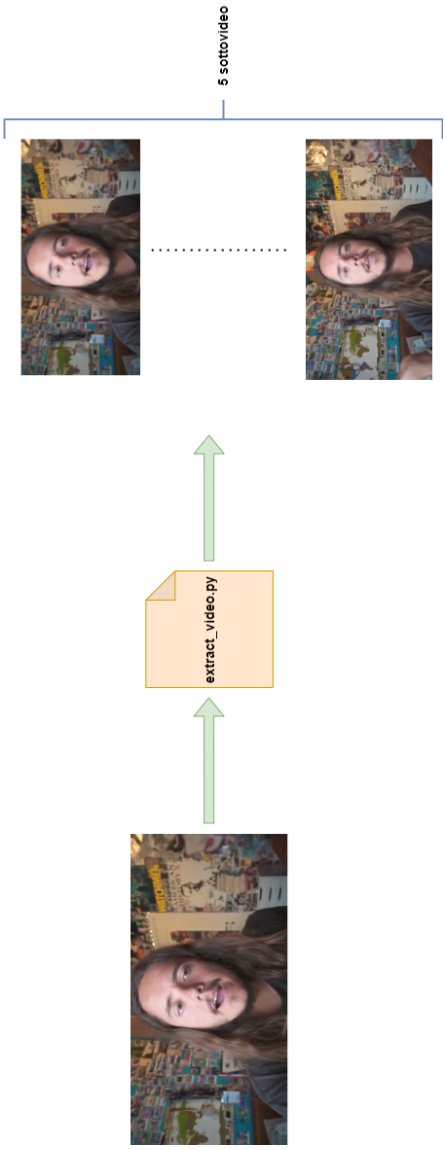

Figura 3.11: Rappresentazione processo dell'algoritmo extract_video.py

Una volta ottenuto un dato complessivo di 1.280 sottovideo divisi allo stesso modo per i video interi, ovvero per lingua in base al sesso e all'età della persona, si passa alla seconda fase principale per il completa-

mento del dataset. Si è utilizzato l'algoritmo *extract_data.py*, scritto anche esso in *Python*, che permette di ottenere la zona labiale per ogni singolo *sottovideo*. Tale algoritmo estrae la zona labiale tramite i landmark che identificano le labbra e il movimento di esse, e di centrarle all'interno della scena del video ridimensionato *300 x 200* pixel, tale taglio permette di escludere qualsiasi altra informazione inutile per il lavoro relativo al viso, come naso, occhi, ecc. Una volta estratto per ogni *sottovideo* la zona labiale del soggetto presente in

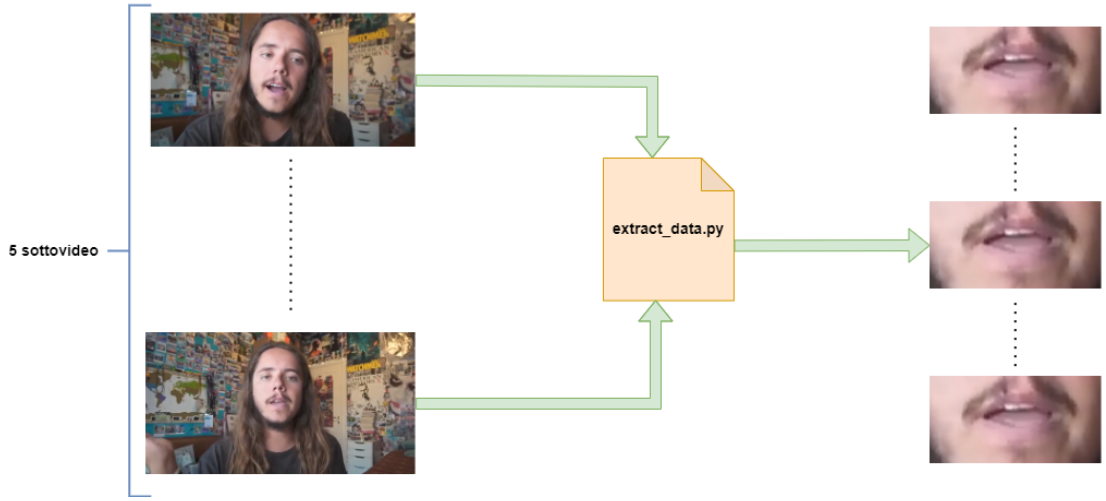

Figura 3.12: Rappresentazione processo dell'algoritmo extract_data.py

esso, il dataset risulta completo e pronta per la fase di *Preprocessing*. Riepilogando, il dataset completo 3.13 conta in totale di 1.280 sequenze video di solo labbra senza audio su 256 soggetti differenti.

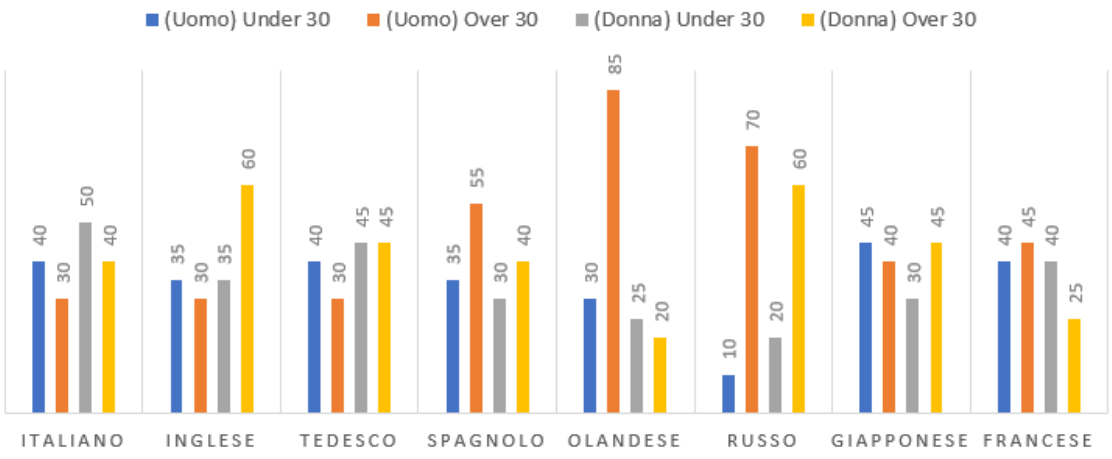

Figura 3.13: Organizzazione del dataset completo

3.3 Preprocessing

Il preprocessing è la fase durante la quale i dati vengono preparati e impacchettati nel formato corretto, in modo tale che il modello scelto possa utilizzarli per l'addestramento e per il testing. Si possono percorrere diverse strategie per rendere compresibili i dati al modello in base alla architettura di quest'ultimo. In questo lavoro si percorre, come già detto, la strada del Deep Learning.

3.3.1 Preprocessing Deep Learning

Il modello utilizzato lavora sulle sequenze video raccolte nella fase precedente, trasformando ogni raccolta di video in un numpy array (con la libreria *Numpy*) e riducendo i dati nel formato (samples, rows, cols, channels) dove *samples* costituisce il numero di video, *rows* e *cols* costituiscono rispettivamente l'altezza e la larghezza di ogni immagine, e *channels* che assume un valore uguale a uno o a tre, indica se le immagini sono in bianco e nero oppure a colori. Successivamente, l'algoritmo procede iniziando a chiedere all'utente che dati deve convertire in dati binari (elementi fondamentali che verranno dati al modello), se dati destinati al training oppure dati destinati al testing. Una volta scelto quali dati convertire, in entrambi i casi, vengono presi in input i video, in formato *.mp4* e dalla durata di 10 secondi ognuno, e vengono estratti i singoli frame come immagini dove verranno convertiti in scala di grigio, in modo tale da alleggerire le prestazioni dell'algoritmo e rendere più stabile la fase di addestramento del modello implementato. Un altro modo per migliorare le prestazioni è quello di effettuare un piccolo ritaglio dell'immagine in modo tale da racchiudere in esso solo la zona labiale per evitare, in casi particolari, oggetti indesiderati che possono confondere il modello in fase di training. Prima del termine dell'algoritmo, verranno inserite le etichette che permettono di essere classificati dal modello, in questo caso verranno inserite la lingua parlata dal soggetto e il suo nome. E infine verranno convertite in dati binari e pronti per essere utilizzati per addestrare e/o testare il modello. Nel Capitolo 4, verranno definiti più nel dettaglio la gestione delle etichette e in che modo viene diviso il dataset in base alle sperimentazioni fatte nel corso di questo lavoro. Come accennato prima, non verranno utilizzate entrambe le tipologie di etichette (lingua parlata e nome) ma verranno usate singolarmente in base allo scopo della sperimentazione. La prima fase del lavoro, che si concentrerà sul riconoscimento della lingua parlata, nella fase di *Preprocessing* verranno inserite le etichette della reale lingua parlata del soggetto. Invece nella seconda fase, si ha come scopo l'identificazione del soggetto, quindi verrà adottata una strategia per la classificazione utilizzando etichette che identificano la reale identità del soggetto.

3.4 Implementazione modello neurale

Prima di descrivere l'implementazione del modello neurale utilizzato per l'identificazione del soggetto tramite la zona labiale, bisogna tenere in considerazione alcune cose. L'implementazione del modello *ConvLSTM2D* è stato preso come riferimento dal lavoro di Emanuele Mezzi [28] che si concentrava sul riconoscimento della lingua parlata tramite le labbra. Per il corretto funzionamento del modello, anche in termini di prestazioni, sono state apportate delle modifiche dall'originale, come ad esempio, l'abilitazione della GPU, per far si che il modello esegue l'addestramento utilizzando la potenza di calcolo offerta dalla scheda video, il numero degli strati e neuroni per ottenere un modello che funzionasse al meglio in base alla potenza della macchina dove sono stati effettuati gli esperimenti. L'implementazione della rete neurale è stata utilizzata la libreria *Keras*, la quale permette di scrivere modelli neurali concentrandosi sugli strati e non sul singolo neurone, e quindi permettendo a colui che implementa di focalizzarsi unicamente sulle caratteristiche del modello utili all'ottenimento della più elevata accuratezza possibile. Si può dividere l'implementazione del modello scelto in quattro fasi:

- **Scelta della strategia:** Deep Learning;
- **Scelta dello strato neurale da utilizzare:** LSTM con l'approccio convoluzionale;
- **Scelta del numero di strati:** il numero di strati può variare a seconda del numero di video con il quale si lavora e/o alla grandezza della *batch_size*;
- **Scelta dei parametri neurali per ciascun strato:** ad esempio, numero di neuroni, funzione di attivazione ed ecc.

3.4.1 Scelta della strategia

Come già anticipato prima, si è scelto la strada del Deep Learning. La differenza, sostanzialmente, dagli altri approcci citati nel Capitolo 2, il Deep Learning si basa anche su sequenze visive, come in questo caso, i video per interi citati nell'organizzazione del dataset e successivamente convertiti in dati binari nella fase di Preprocessing.

3.4.2 Scelta dello strato neurale da utilizzare

Per la strategia scelta, descritta prima, si è voluto utilizzare ciò che metteva a disposizione la libreria *Keras* per il Deep Learning, ovvero gli strati LSTM di tipo convoluzionale. Di base si è utilizzato l'elaborazione del flusso dei dati in modo monodirezionale e nel caso se si preferisce un approccio bidirezionale, l'unica modifica da apportare sarebbe l'aggiunta di una funzione apposita, che permetta di ottenere questa ulteriore caratteristica.

3.4.3 Scelta del numero di strati

Per quanto riguarda il numero di strati utilizzati nell'implementazione del modello, non è stato effettuato un calcolo preciso che permetteva di ottenere il risultato ottimale. Ma si sono svolti vari tentativi che permettevano il corretto funzionamento e stabilità del modello basandosi sulle performance della macchina che si aveva a disposizione e sulla quantità di dati che il modello doveva elaborare. Nel Capitolo 4, dove verranno

descritti le varie sperimentazioni, ci sono dei piccolissimi cambiamenti riguardo al numero di strati utilizzati quando verranno effettuati gli addestramenti al modello con *batch_size* uguale a 6 perchè quest'ultimo non si addestrava correttamente e necessitava di una potenza di calcolo maggiore.

3.4.4 Scelta dei parametri neurali per ciascun strato

Anche in questo caso sono stati svolti vari tentativi per ottenere un corretto funzionamento del modello in base alla potenza di calcolo che si aveva a disposizione. A partire dal numero di neuroni che nelle varie sperimentazioni ha avuto un leggero cambiamento per quanto si voleva effettuare sperimentazioni con *batch_size* uguale a 6 o maggiore. Sono state scelte le adeguate funzioni di attivazioni che permettevano di avere una giusta stabilità del modello durante la fase di training e di testing ed un'ottimale accuratezza di esso.

Capitolo 4

Sperimentazioni e risultati

Dopo aver definito nel capitolo precedente le strategie utilizzate per l'estrazione della zona labiale e il modello neurale utilizzato per l'addestramento, in questo capitolo verranno descritti le sperimentazioni svolte in questo lavoro con i propri risultati. Le seguenti sperimentazioni si possono dividere in tre fasi:

- *Riconoscimento della lingua parlata*
- *Identificazione del soggetto*
- *Identificazione del soggetto con l'integrazione della lingua parlata*

Tali sperimentazioni sono state effettuate tutte con la medesima configurazione del modello, cambiando solamente le etichette ai dati binari in modo da classificare i video in base al tipo di sperimentazione.

4.1 Gestione delle etichette

La prima e seconda sperimentazione differisco nel modo in cui vengono classificati i video. Per la prima sperimentazione i video presi in esame vengono etichettati con un numero che va da 0 a 7 in fase di preprocessing. Il numero corrisponde alla lingua parlata dal soggetto presente nel video permettendo al modello, in fase di valutazione e/o di training, di classificare i video in base alla loro lingua originale.

Lingua	**Etichetta**
Italiano	0
Inglese	1
Tedesco	2
Spagnolo	3
Olandese	4
Russo	5
Giapponese	6
Francese	7

Tabella 4.1: Legenda delle etichette usate nel *"Riconoscimento della lingua parlata"*

Nella seconda sperimentazione, invece, è stata utilizzata la *One Hot Encoding* per facilitare al modello la classificazione dei video. La motivazione di tale scelta è dovuta al fatto che avendo un numero di soggetti molto elevato, ovvero 256, etichettarli con un numero che andava da 0 a 255 non sembrava adatto al modello. Nello specifico, il funzionamento della codifica *One Hot* verrà descritto successivamente nella sezione "*Identificazione del soggetto*" di questo capitolo.

Nome video	**Etichetta**	**One Hot Encoding**
1_1_1_1	1111	[1 0 ... 0 ... 0]
1_1_1_2	1112	[0 1 ... 0 ... 0]
...	...	...
...	...	...
...	...	...
...	...	...
8_1_1_1	8111	[0 0 ... 0 ... 1 0]
8_1_1_2	8112	[0 0 ... 0 ... 0 1]

Tabella 4.2: Legenda delle etichette per l'*"Identificazione del soggetto"*

4.2 Configurazione modello

Prima di andare a descrivere le varie sperimentazioni, bisogna tenere in considerazione un fattore molto importante. I risultati prodotti da tutte le sperimentazioni, poco prima citate, sono stati ottenuti dalla stessa configurazione del modello. Prima delle sperimentazioni sono stati svolti dei tentativi di configurazioni ideale per far si che l'addestramento andasse a buon fine tenendo in considerazione la potenza di calcolo della macchina che si aveva a disposizione. La configurazione utilizzata in tutte le sperimentazioni è la seguente:

- **Numero di strati:** sono stati utilizzati 2 strati ricorrenti, 1 *flatten* per rendere l'output monodimensionale e 2 strati densi.
- **Tipi di strati utilizzati:** *ConvLSTM2D, Flatten, Dense*.
- **Numero di celle neurale utilizzate per ogni strato:** il primo strato ricorrente contiene 64 celle neurali, mentre per il secondo 96. Il primo strato *denso* ne contiene 16 e, infine, per il secondo ha 8 celle neurali che indica il numero di lingue utilizzate per la classificazione 4.2. Per la sperimentazione dell'identificazione invece, in quest'ultimo strato denso verranno messe ben 256 celle neurali che indicheranno il numero di soggetti distinti presenti nel dataset 4.1.
- **Funzioni di attivazione utilizzate:** per gli strati ricorrenti è stata utilizzata la funzione tangente, per il primo strato denso nessuna funzione, mentre per lo strato di output la funzione softmax.
- **Ottimizzatore utilizzato:** *Adamax*.
- **Learning rate:** *0.0001*
- **Batch size utilizzata:** 6
- **Epoche utilizzate:** 200 epoche
- **Step per epoca:** 32 steps per epoca

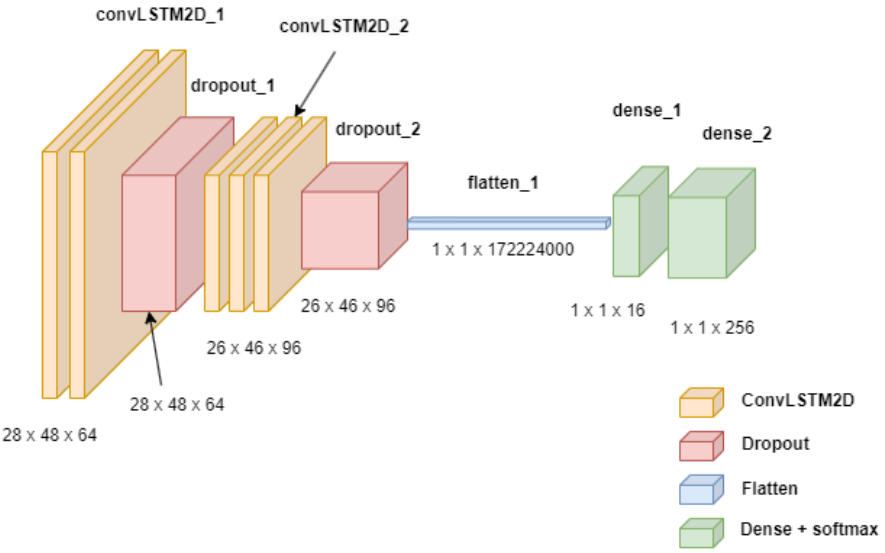

Figura 4.1: Struttura del modello utilizzato per l'*"Identificazione del soggetto"*

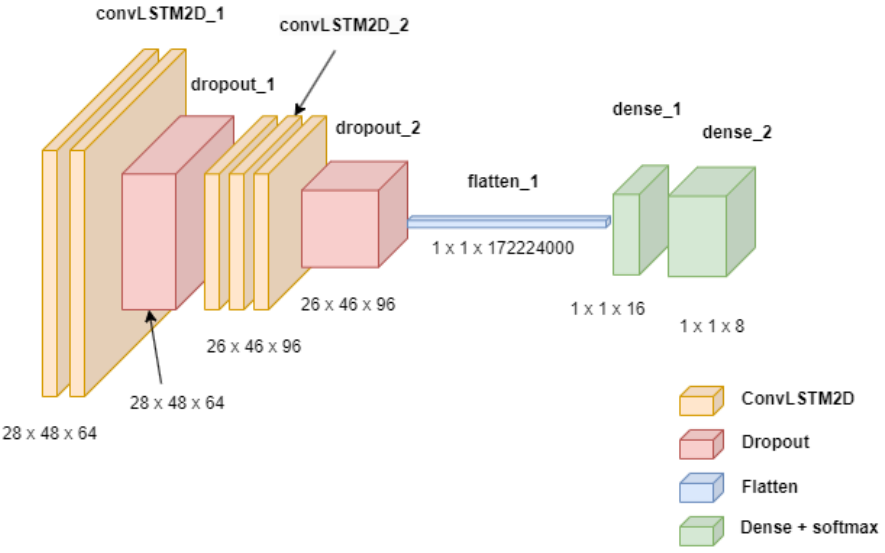

Figura 4.2: Struttura del modello utilizzato per il *"Riconoscimento della lingua parlata"*

Riassumendo, si è utilizzato *softmax* come funzione di attivazione nell'ultimo livello, l'ottimizzatore Adam con un learning rate di *0.0001*. La loss è stata calcolata in termini di *sparse categorical cross-entropy* per la sperimentazione del *"Riconoscimento della lingua parlata"*, invece per l'*"Identificazione del soggetto"* è stata calcolata utilizzando la funzione di *categorical cross-entropy* la cui motivazione è dovuta al fatto che le etichette utilizzate per l'identificazione andavano in conflitto con l'addestramento del modello. Il modello è stato addestrato per 200 epoche con una dimensione di batch size pari a 6.
Infine, si precisa che i risultati ottenuti nel corso di questo lavoro saranno inferiori rispetti a quelli citati nei lavori correlati di entrambe le problematiche che affronta questo lavoro, nel Capitolo 2, tale motivazione è data dal tipo di approccio che si è utilizzato, ossia il Deep Learning. Questa strategia comporta un'allocazione di memoria molto elevata e nell'implementazione ha comportato una riduzione minimale degli strati neurali e delle celle neurali contenute in essi, portando alla configurazione appena descritta.

4.3 Riconoscimento della lingua parlata

Questa sperimentazione si basa principalmente sul riconoscimento della lingua parlata da un soggetto presente nel video e viene effettuata in due modi diversi:

- *Riconoscimento della lingua parlata di un soggetto visto*
- *Riconoscimento della lingua parlata di un soggetto non visto*

La differenza, sostanzialmente, è come il dataset viene diviso per l'addestramento e la valutazione del modello neurale.

4.3.1 Riconoscimento della lingua parlata di un soggetto visto

In questa sperimentazione si è diviso il dataset in modo tale da permettere al modello di riconoscere la lingua di un soggetto che lui ha già precedentemente visto. Avendo descritto nel Capitolo 3 l'organizzazione del dataset, si ottiene che ogni soggetto abbia un numero di 5 video dalla durata di 10 secondi. Quindi la divisione del dataset è stata effettuata in modo tale da avere 4 video come dati di training e 1 video come dato di testing per ogni soggetto, la selezione di tali video vengono presi totalmente in modo casuale tenendo in considerazione che il totale di video destinati per il testing saranno *256* e per il training *1024*. E' stata presa in considerazione tale scelta della divisione 4.3 del dataset perchè si avvicina più all'idea dell'identificazione del soggetto.

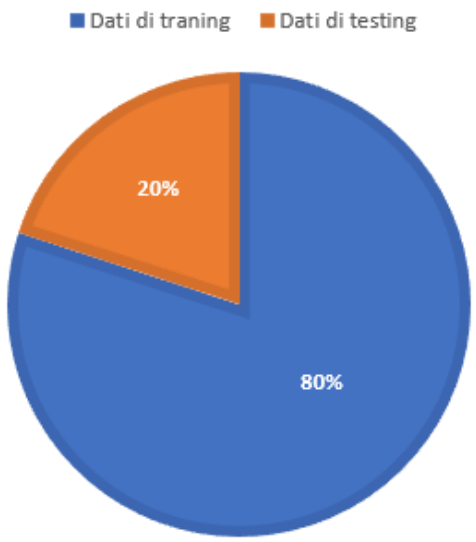

Figura 4.3: Divisione del dataset per la sperimentazione di riconoscimento della lingua con soggetto già visto

Il modello ha dato come risultato migliore il 74,21% di accuratezza e successivamente saranno raffigurati il grafico di andamento del training e della validation del modello, dove quest'ultima conterrà dati di testing, e la matrice di confusione dove verranno visualizzate le risposte date dal modello 4.4.

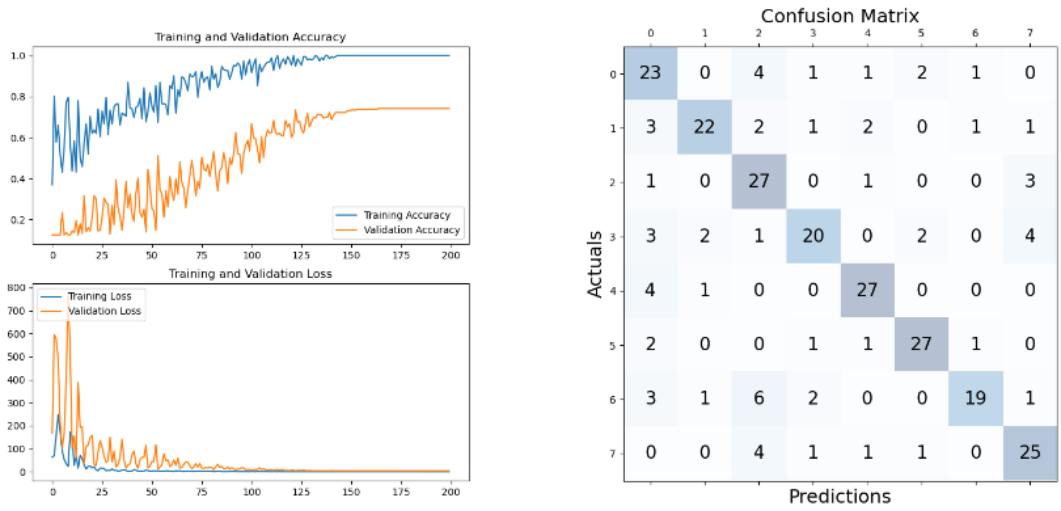

Figura 4.4: Valutazione del modello per il riconoscimento della lingua parlata con il soggetto visto

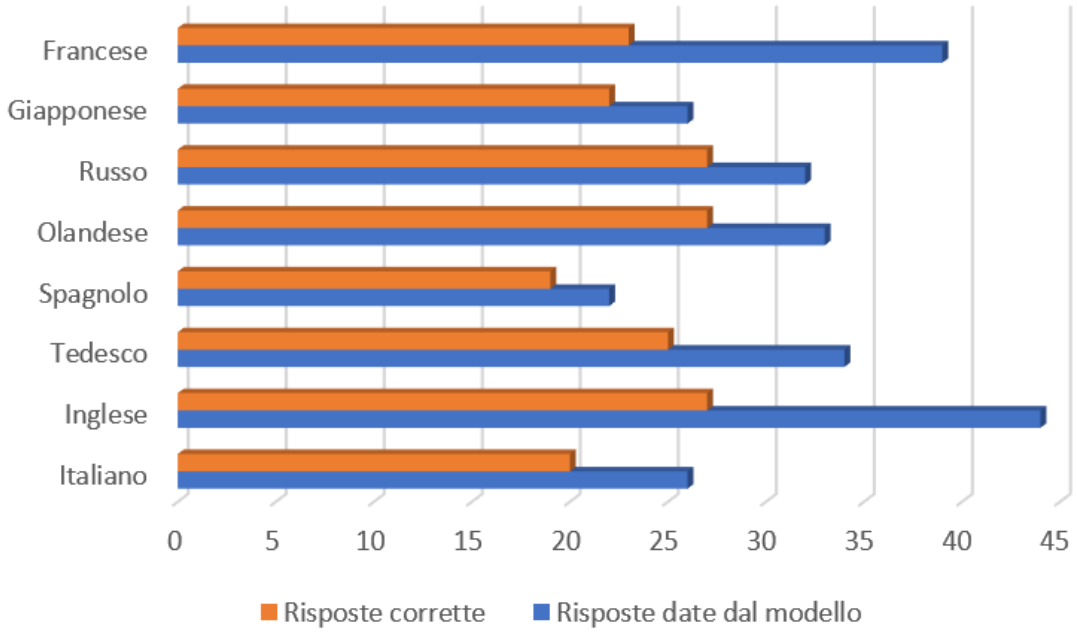

Figura 4.5: Analisi delle risposte date dal modello per il riconoscimento della lingua parlata con soggetto visto

Come si può vedere, la matrice di confusione permette di analizzare a fondo le risposte che sono state fornite dal modello e quali tra queste sono risultate esatte. Per facilitare la comprensione della matrice di confusione si segua questa legenda:

Lingua	**Indice**
Francese	0
Giapponese	1
Inglese	2
Italiano	3
Olandese	4
Russo	5
Spagnolo	6
Tedesco	7

Si può notare che, nonostante si abbia ottenuto un discreto risultato per quanto riguarda le risposte corrette, il modello ha presentato delle difficoltà nel riconoscere alcune lingue. Come ad esempio, *francese-inglese*, *italiano-tedesco*, *olandese-francese*, *spagnolo-inglese* e *tedesco-spagnolo*.

4.3.2 Riconoscimento della lingua parlata di un soggetto non visto

La differenza con la sperimentazione precedentemente descritta è il modo in cui viene diviso il dataset. Il dataset è stato diviso seguendo il concetto di *"Subject Independet"*, ovvero si è pensato di dare al modello soggetti dove egli non ha visto in fase di training. La divisione è leggermente diverso da quella descritta nella sperimentazione precedente, infatti verranno considerati i soggetti. Nel senso che verranno presi, in maniera del tutto casuale, 4 soggetti per lingua destinati come dati di *testing*, altri 4 soggetti come dati di *validation* e i restanti come dati di *training*. Il dataset risulta diviso nel seguente grafico:

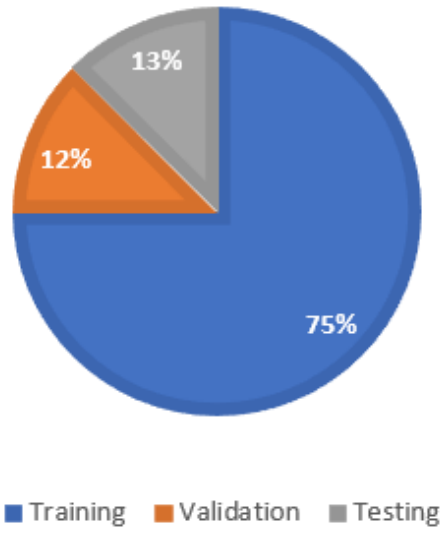

Figura 4.6: Divisione del dataset per la sperimentazione di riconoscimento della lingua con soggetto non visto

Il modello, con la stessa configurazione citata prima, ha dato come risultato migliore il 36,78% di accuratezza. Tale risultato risulta così basso rispetto quello dato dalla sperimentazione precedente è dovuto al fatto che il dataset, organizzato nel Capitolo 3, è abbastanza piccolo per questa sperimentazione e che il modello non ha mai visto i soggetti destinati nel testing nella fase di training. Tuttavia, nonostante ottenendo un'accuratezza molto bassa rispetto alla sperimentazione precedente, comunque si è ottenuto un risultato superiore al lavoro di riferimento di Emanuele Mezzi [28] che lui stesso definiva tale sperimentazione come *"Subject Independent"*. Anche in questo caso sono stati raffigurati i grafici riguardo al risultato ottenuto 4.7.

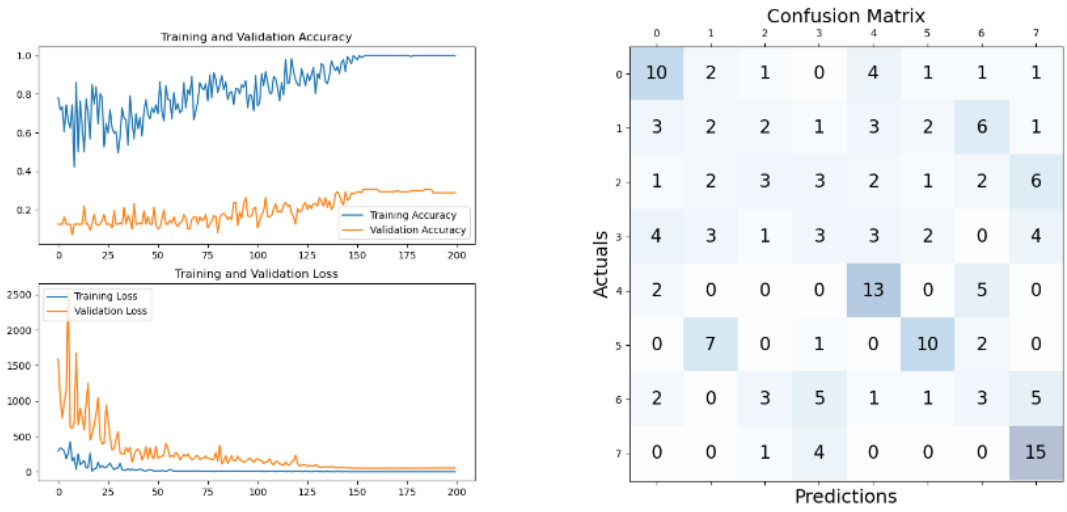

Figura 4.7: Valutazione del modello per il riconoscimento della lingua parlata con il soggetto non visto

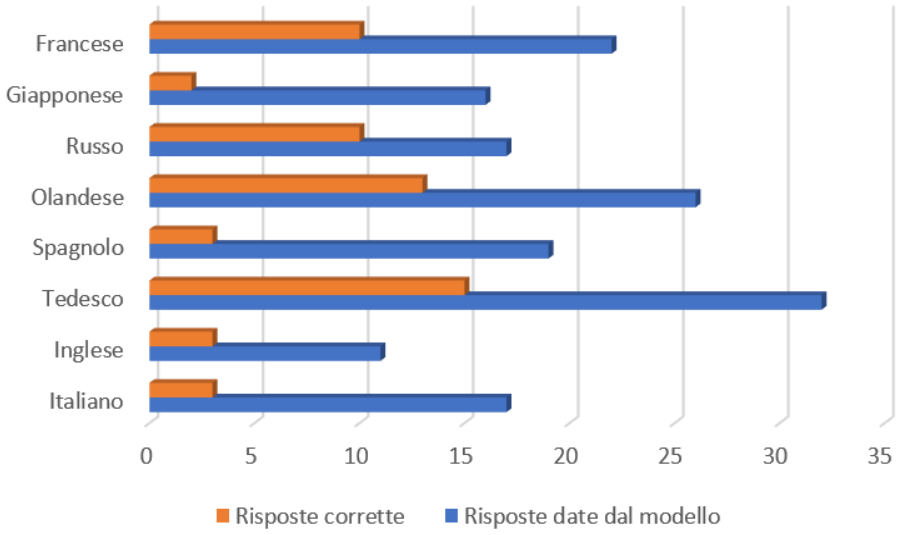

Figura 4.8: Analisi delle risposte date dal modello per il riconoscimento della lingua parlata con soggetto non visto

Nella seguente valutazione la matrice di confusione è completamente differente in confronto a quella raffigurata nella sperimentazione *"Riconoscimento della lingua parlata con soggetto visto"*. Tenendo in considerazione che per ogni lingua sono stati utilizzati 4 soggetti diversi per un totale di 20 video, si può notare che le lingue che hanno riscontrato un esito "positivo" sono le seguenti: *Francese, Olandese, Russo* e *Tedesco*. In confronto alla matrice di confusione, analizzata nella precedente sperimentazione, le lingue che hanno causato problemi al modello sono diverse. Nella seguente matrice 4.7 si possono notare le seguenti problematiche: *francese-olandese*, *giapponese-spagnolo*, *inglese-tedesco*, *italiano-francese/tedesco*, *olandese-spagnolo*, *russo-giapponese*, *spagnolo-italiano/tedesco* e infine *tedesco-italiano*.

4.4 Considerazioni sui risultati del riconoscimento della lingua parlata

Dalle matrici di confusione, nonostante le evidenti differenze tra la sperimentazione in cui viene isolata la componente biometrica e non, si confermano comunque le seguenti affermazioni. In entrambe le figure, infatti, si può osservare come il modello, a prescindere dall'identità nota o meno del soggetto, tenda a non confondere la lingua russa con quella inglese. La pronuncia dell'inglese è molto diversa da quella del russo [29], poiché l'inglese ha molte più vocali aperte e chiuse rispetto al russo. Inoltre, l'inglese ha molte più vocali dure, come "a" e "e", mentre il russo ha molte più vocali morbide, come "o" e "e". Il russo inoltre utilizza le labbra per produrre suoni gutturali come "g" o "k", mentre l'inglese non ha suoni gutturali simili. Inoltre, il russo ha una serie di suoni labiali sibilanti che sono prodotti con la posizione delle labbra. Questi suoni non esistono in inglese. In inglese, l'articolazione labiale è meno evidente e le variazioni nella posizione delle labbra durante la produzione di un suono sono meno pronunciate. Queste differenze rendono i due idiomi molto diversi l'uno dall'altro.

4.5 Identificazione del soggetto

A differenza delle sperimentazioni già definite precedentemente, lo scopo di questa corrente è riuscire a riconoscere l'identità di un soggetto utilizzando, come sempre, la zona labiale. Come espresso precedentemente, in questa sperimentazione la divisione del dataset 4.3 è uguale a quella citata nel paragrafo riguardo al *riconoscimento della lingua parlata di un soggetto visto*. Ovvero, sapendo che ogni soggetto ha 5 video dalla durata di 10 secondi vengono presi 4 come dati di training e 1 video viene preso come dato di testing per ogni soggetto. Lo scopo di questa sperimentazione è simile a quella citata precedentemente, ovvero il modello deve riconoscere l'identità di una persona già vista in fase di training. La differenza è che i video, in fase di preprocessing, vengono etichettati in base ad una serie di numeri presi dalla nomenclatura del video. Come espresso nel Capitolo 3, la nomenclatura di ogni singolo video è composta dal seguente formato, ossia *"(lingua)_(sesso)_(età)_(id unico)_(numero di framerate)"*. Successivamente quando vengono estratti i sottovideo per ogni soggetto viene aggiunto un ulteriore numero che rappresenta la parte del video, *"(lingua)_(sesso)_(età)_(id unico)_(numero di framerate)_**(numero parte)**"*. In fase di preprocessing, viene preso il numero che identifica la lingua, il sesso, l'età e l'id unico del video eliminando il carattere "_" e successivamente messi insieme ottenendo un numero identificativo per il singolo soggetto (ad esempio, 8_1_1_1 = 8111). Una volta ottenuto il numero identificativo del soggetto viene convertito in un array tramite *One Hot Encoding* per dare la possibilità al modello di classificare i video di training e di testing senza problemi. L'array 4.9 dato dalla codifica One Hot è semplicemente una serie di bit, che corrisponde al numero di elementi presenti per la classificazione (in questo caso 256), dove tutti i bit assumono valore 0 tranne per un singolo bit che sarà uguale a 1, permettendo di identificare a quale oggetto si riferisce.

Figura 4.9: Elaborazione dell'etichetta per la sperimentazione di identificazione in fase di preprocessing

Dopo questo chiarimento riguardo alla gestione delle etichette in questa sperimentazione, il modello implementato ha prodotto come risultato il 49,60% di accuratezza. Non è stato possibile raffigurare la matrice di confusione perchè, come si può intuire, risultava molto difficile da comprendere per via del fatto del elevato numero di soggetti da classificare, ovvero 256. In compenso, si può vedere il numero totale di risposte corrette date dal modello nella Figura 4.11.

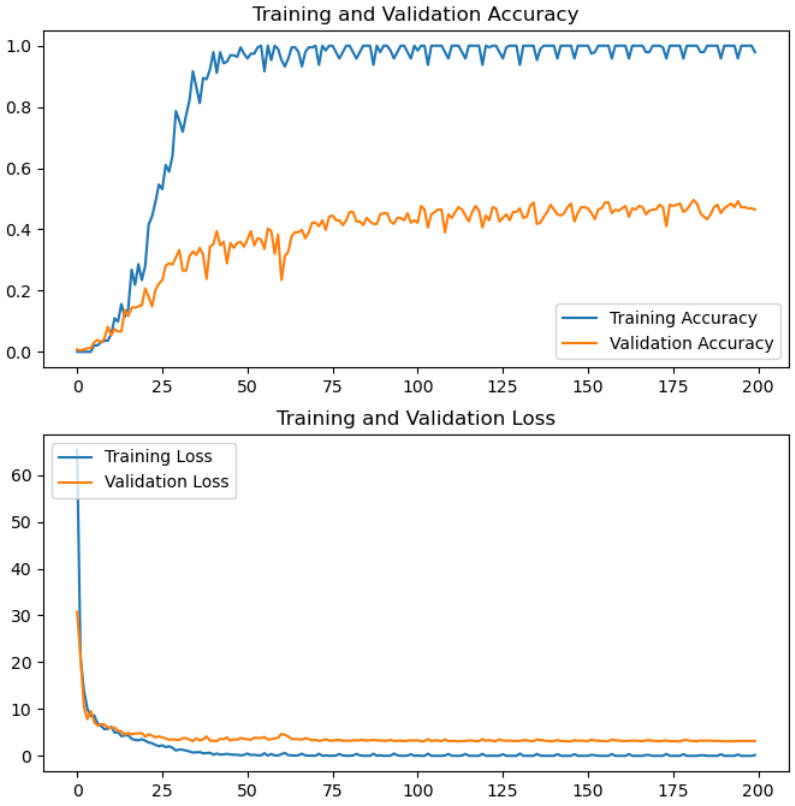

Figura 4.10: Andamento di training della sperimentazione dell'identificazione

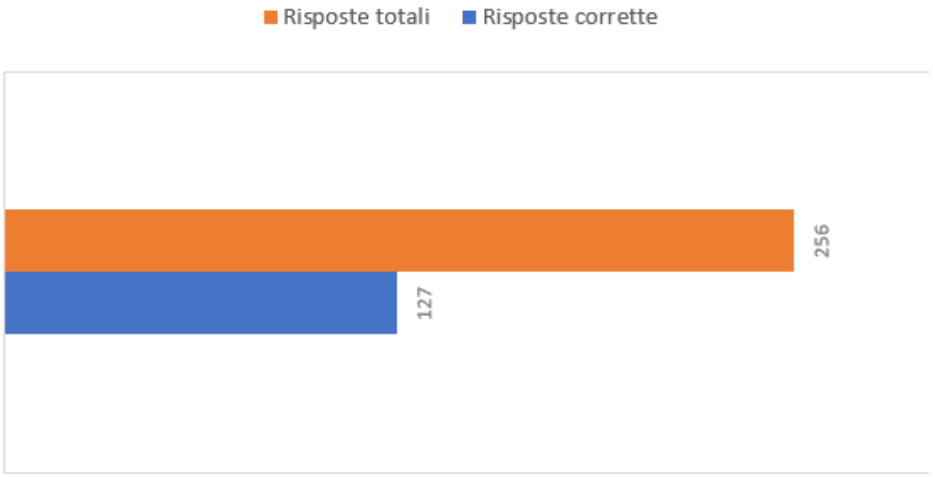

Figura 4.11: Valutazioni risposte sperimentazione dell'identificazione

4.6 Identificazione del soggetto con l'integrazione della lingua parlata

In quest'ultima fase di lavoro verranno considerati i risultati ottenuti dalle seguenti sperimentazioni: *Riconoscimento della lingua parlata con soggetto visto* e *Identificazione del soggetto*. Lo scopo stabilito in questo lavoro è quello di migliorare l'identificazione di un soggetto integrando la lingua parlata da esso, e per realizzarlo è stato deciso di fondere insieme i risultati ottenuti dalle sperimentazioni citate poco prima. Riepilogando, al di sotto verranno raffigurati i risultati ottenuti fino ad adesso:

	Riconoscimento della lingua parlata		**Identificazione del soggetto**
	Soggetto visto	**Soggetto non visto**	
Accuracy (%)	74,21	36,78	49,60

Tabella 4.3: Tabella riassuntiva dei risultati ottenuti

La motivazione per cui si è sfruttato il risultato migliore dato dalla sperimentazione *"Riconoscimento della lingua parlata con soggetto visto"* è data al fatto che la strategia utilizzata per la divisione del dataset è più coerente ed adatta allo scopo dell'identificazione, ovvero il modello visualizza il soggetto in fase di training e lo ritrova in fase di testing.

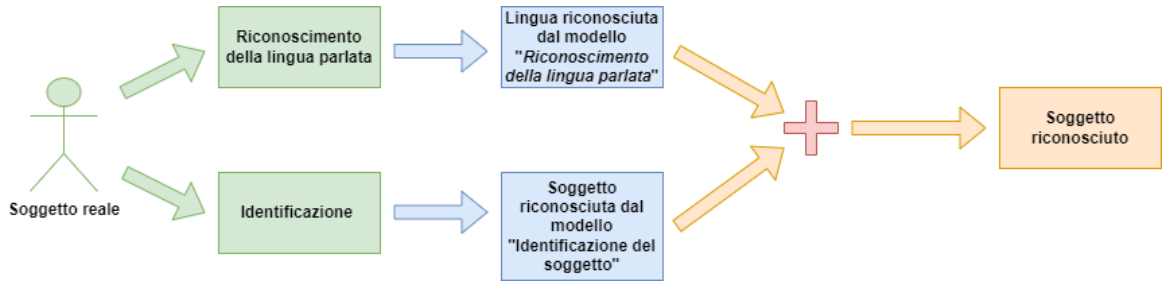

Figura 4.12: Raffigurazione del processo di fusione dei risultati del "*Riconoscimento lingua con soggetto visto*" e "*Identificazione del soggetto*"

Il modo in cui sono stati fusi i risultati dei due modelli è il seguente:

1. Nella fase di valutazione della sperimentazione "*Identificazione del soggetto*" vengono prese per ogni predizione 8 valori più alti. Una singola predizione, data da un video contenuto nella cartella di testing, offerto al modello con il metodo "*predict*" (fornita dalla libreria *Keras*) produrrà un array di 256 elementi (che corrisponde al numero di identità distinte presente nel dataset). Questo array conterrà numeri che assumeranno un valore compreso tra 0 e 1, ed il numero che ha il valore più alto corrisponde ad un determinato soggetto dove il modello crede di sapere chi esso sia. Complessivamente, da come è stato descritta la divisione del dataset 4.3 per questa sperimentazione, si otterranno 256 predizioni dalla dimensione di 256 valori ognuno. Quindi, per ogni predizione, vengono presi 8 soggetti con il valore più alto di predizione e stampati in un file *.csv* per rendere facile la comprensione di esso. Per identificare le 8 predizioni per ogni dato di testing, viene preso in considerazione il video di riferimento. La scelta di prendere 8 soggetti con la predizione più alta è dovuta al fatto che si voleva tenere lo stesso numero delle lingue utilizzate per questo lavoro.

2. Una volta ottenuto le informazioni appena descritte, si prosegue con l'integrazione della lingua parlata nella valutazione dell'identificazione del soggetto. Come sono stati descritti precedentemente, sono noti le seguenti affermazioni:

    - Nel corrente capitolo nella sezione dove si descrive la sperimentazione *"Riconoscimento della lingua parlata"*, i video vengono etichettati con un numero che assume valore da 0 a 7 (4.1), dove tale numero identificherà la lingua parlata dal soggetto presente nel video.
    - Nel Capitolo 3, precisamente nella sezione *"Creazione del dataset"*, viene indicata la nomenclatura fornita per ogni video. Il formato di tale nomenclatura è la seguente: *(lingua_sesso_età_id-unico_framerate)*.

    Queste affermazioni appena citate saranno utili per la comprensione del seguente passaggio. Nella fase di valutazione sono stati presi in considerazione, come si è accennato prima, i due modelli che hanno offerto il risultato migliore descritto nelle sperimentazioni precedenti. Per fare chiarezza, sono stati presi quelli inerenti al *"Riconoscimento della lingua parlata con soggetto visto"* e *"Identificazione del soggetto"*. La prima parte della valutazione si occupa di identificare il soggetto ottenendo l'etichetta associata alla persona che il modello predice, ad esempio, l'etichetta *8111*. Successivamente entra in gioco la valutazione dello stesso soggetto, identificato secondo il modello di *identificazione*, per il riconoscimento della lingua parlata. Nel seguente caso si sfrutta la nomenclatura dell'etichetta ottenuta nell'identificazione, definita precedentemente. Infatti è noto che il primo carattere dell'etichetta corrisponde alla lingua parlata dal soggetto. Producendo la predizione della lingua, offerta dal modello del *"Riconoscimento della lingua parlata con soggetto visto"*, si otterrà un'etichetta dove può assumere un valore da 0 a 7. Bisogna considerare che la numerazione delle lingue utilizzate nelle etichette dell'identificazione vanno da 1 fino a 8, invece quelle delle lingue da 0 a 7, quindi si può dedurre che l'etichetta della lingua predetta dal modello deve essere aggiunta un'unità per effettuare il giusto confronto tra le etichette usate nell'identificazione. Ottenuto la predizione della lingua del medesimo soggetto, prodotto dal modello di identificazione, viene utilizza l'etichetta predetta della lingua e viene confrontata tra i primi caratteri di tutte le etichette di identificazioni presenti nel file *.csv* contenente gli 8 soggetti con predizione più alta riferiti allo stesso video utilizzato sia per la valutazione del riconoscimento della lingua che per l'identificazione. Se è presente almeno un'etichetta che inizia con lo stesso numero prodotto dal riconoscimento della lingua viene presa quella con la predizione più alta e viene data come risposta la sua etichetta per intero, invece se non è presente viene riconfermata la risposta data dal modello di identificazione.

L'idea di base è quella appena descritta, ma è stata aggiunta un piccolo "*check*" per ottenere un miglioramento significativo. La motivazione di tale aggiunta è dovuta dal fatto che in molti casi il modello per l'identificazione riusciva a scovare la reale identità di un soggetto ma con l'integrazione della lingua parlata il risultato definitivo veniva distorto da quest'ultima. E' stato idealizzato una certa soglia di predizione dove l'integrazione dei risultati del "*Riconoscimento della lingua parlata*" poteva avvenire se tale predizione fosse stata più piccola del limite impostato. Dopo una lunga analisi, la soglia viene impostata a *0.45* perchè ispezionando i risultati dati dall'identificazione la maggior parte di essi scovavano l'identità reale con predizione maggiore di *0.45*, quindi con una predizione inferiore viene concessa l'integrazione della lingua parlata.
La strategia appena formulata per la fusione dell'"*Identificazione del soggetto*" e il *"Riconoscimento della lin-*

gua parlata con soggetto visto" ha prodotto un miglioramento del 6% nell'identificazione biometrica tramite zona labiale. Basta pensare che i risultati raffigurati in Figura 4.11, in quest'ultima sperimentazione sono stati ottenuti **142** risposte corrette date dalla fusione dei due modelli, e infine producendo un'accuratezza del *55,46%*.

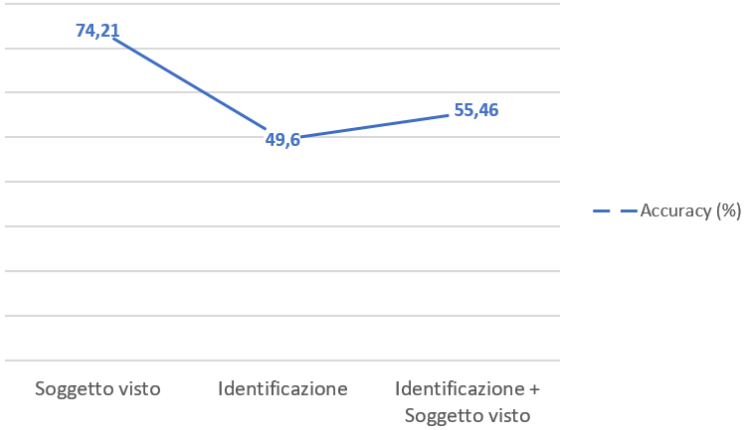

Figura 4.13: Miglioramento dell'identificazione con integrazione della lingua parlata

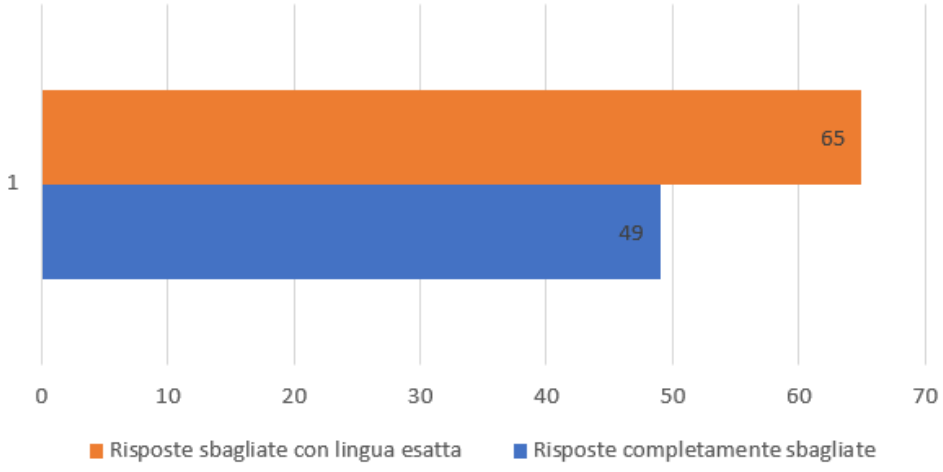

Figura 4.14: Analisi risposte sbagliate con lingua predetta corretta

Prima di concludere, nel il grafico raffigurato precedentemente 4.14 si può notare un'analisi riguardo alle risposte sbagliate date dal modello con l'integrazione della lingua parlata. Come si può intuire, le risposte sbagliate predette dal modello integrando la lingua risultano essere in totale *114*. Tuttavia, bisogna precisare che le *65* risposte sbagliate, raffigurate in figura, sono state causate perchè, nonostante la lingua predetta fosse corretta integrandola all'identificazione non si aveva la possibilità di cambiare la predizione dell'identità per via del fatto che all'interno della stima degli 8 soggetti non conteneva il soggetto predetto che parlasse quella lingua.

Capitolo 5

Conclusioni

Nel seguente lavoro appena affrontato nei capitoli precedenti si poneva come obiettivo un miglioramento nell'identificazione biometrica tramite la zona labiale integrando la lingua parlata del soggetto, infatti si è ottenuto un miglioramento del *6%*. L'approccio scelto, per ottenere tali risultati, è stato l'utilizzo del Deep Learning che ha consentito di lavorare direttamente con sequenze visive escludendo completamente l'audio. Tale scelta ha comportato riduzioni di strati e celle neurali per poter affrontare le sperimentazioni svolte in modo tale da occupare una sufficiente allocazione di memoria per ottenere risultati stabili e coerenti. Difatti dal punto di vista computazionale, si possono trovare altre strategie per ottimizzare al meglio l'allocazione di memoria da poter incrementare l'efficacia del modello neurale utilizzato in questo lavoro. In più, sarà sicuramente possibile migliorare i risultati raggiunti, in quanto questo lavoro è stato realizzato basandosi unicamente su otto lingue e questo, chiaramente, comporta ad una valutazione limitata. In futuro, il data-set elaborato può essere migliorato comprendendo più lingue, e soprattutto inserendo più soggetti diversi per permettere al modello implementato di lavorare su più dati ed ottenere un risultato più coerente. Infine, l'efficacia di questo lavoro è rendere l'identificazione biometrica più consistente e difficile da replicare, soprattutto per i malintenzionati, difatti l'obiettivo di questo operato è ottenere una visione più realistica sulla fortificazione di un qualsiasi sistema informatico in qualunque ambito di lavoro o non.

Bibliografia

[1] N. A. Thacker J. Luettin e S. W. Beet. "Speaker identification by lipreading". In: *Proc. 4th Int. Conf. ICSLP* (1996), pp. 62–65.

[2] D. Thambiratnam T. Wark e S. Sridharan. "Person authentication using lip information". In: *Proc. TENCON Brisbane - Australia. IEEE TENCON Region 10 Annu. Conf.* Speech Image Technol. Comput. Telecommun. (1997), pp. 153–156.

[3] U. Dieckmann R. W. Frischholz. "BioID: a multimodal biometric identification system". In: *Computer 33* 2.1 (2000), pp. 64–68.

[4] Feng Cheng, Shi-Lin Wang e Alan Wee-Chung Liew. "Visual speaker authentication with random prompt texts by a dual-task CNN framework". In: *Pattern Recognition* 83 (2018), pp. 340–352. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2018.06.005. URL: https://www.sciencedirect.com/science/article/pii/S0031320318302152.

[5] T. Wark, S. Sridharan e V. Chandran. "The use of temporal speech and lip information for multimodal speaker identification via multi-stream HMMs". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. Vol. 4. 2000, 2389–2392 vol.4. DOI: 10.1109/ICASSP.2000.859322.

[6] Shi-Lin Wang e Alan Wee-Chung Liew. "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power". In: *Pattern Recognition* 45.9 (2012). Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), pp. 3328–3335. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.02.016. URL: https://www.sciencedirect.com/science/article/pii/S0031320312000787.

[7] Xin Liu e Yiu-ming Cheung. "Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification". In: *IEEE Transactions on Information Forensics and Security* 9.2 (2014), pp. 233–246. DOI: 10.1109/TIFS.2013.2293025.

[8] J. Luettin, N.A. Thacker e S.W. Beet. "Speaker identification by lipreading". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. Vol. 1. 1996, 62–65 vol.1. DOI: 10.1109/ICSLP.1996.607030.

[9] H.E. Cetingul et al. "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading". In: *IEEE Transactions on Image Processing* 15.10 (2006), pp. 2879–2891. DOI: 10.1109/TIP.2006.877528.

[10] Simon Lucey. "An Evaluation of Visual Speech Features for the Tasks of Speech and Speaker Recognition". In: *Audio- and Video-Based Biometric Person Authentication*. A cura di Josef Kittler e Mark S. Nixon. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 260–267. ISBN: 978-3-540-44887-7.

[11] Chi Ho Chan et al. "Local Ordinal Contrast Pattern Histograms for Spatiotemporal, Lip-Based Speaker Authentication". In: *IEEE Transactions on Information Forensics and Security* 7.2 (2012), pp. 602–612. DOI: 10.1109/TIFS.2011.2175920.

[12] K. Messer et al. "Xm2vtsdb: The extended m2vts database". In: *Proc. of Audio- and Video-Based Person Authentication* (apr. 2000).

[13] Carlos M. Travieso et al. "Using a Discrete Hidden Markov Model Kernel for lip-based biometric identification". In: *Image and Vision Computing* 32.12 (2014), pp. 1080–1089. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2014.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0262885614001474.

[14] M.J. Jones P. Viola. "Robust Real-Time Face Detection". In: *International Journal of Computer Vision* 57.12 (2004), pp. 137–154.

[15] Yujie Dong e Damon L. Woodard. "Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study". In: *2011 International Joint Conference on Biometrics (IJCB)*. 2011, pp. 1–8. DOI: 10.1109/IJCB.2011.6117511.

[16] "PIE Face Database, Available: http://www.ri.cmu.edu/research_project_detail.html?" In: ().

[17] Oliver Langner et al. "Presentation and validation of the Radboud Faces Database". In: *Cognition and Emotion* 24.8 (2010), pp. 1377–1388. DOI: 10.1080/02699930903485076. eprint: https://doi.org/10.1080/02699930903485076. URL: https://doi.org/10.1080/02699930903485076.

[18] Meng Liu et al. "DeepLip: A Benchmark for Deep Learning-Based Audio-Visual Lip Biometrics". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2021, pp. 122–129. DOI: 10.1109/ASRU51503.2021.9688240.

[19] Carrie Wright e Darryl William Stewart. "Understanding visual lip-based biometric authentication for mobile devices". In: *EURASIP Journal on Information Security volume* 3 (2020). DOI: 10.1186/s13635-020-0102-6.

[20] Carrie Wright e Darryl Stewart. "One-Shot-Learning for Visual Lip-Based Biometric Authentication". In: *Advances in Visual Computing*. A cura di George Bebis et al. Cham: Springer International Publishing, 2019, pp. 405–417. ISBN: 978-3-030-33720-9.

[21] Krzysztof Wrobel et al. "Using a Probabilistic Neural Network for lip-based biometric verification". In: *Engineering Applications of Artificial Intelligence* 64 (2017), pp. 112–127. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2017.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0952197617301227.

[22] J. Wang et al. "A Large-Scale Depth-Based Multimodal Audio-Visual Corpus in Mandarin". In: *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. Los Alamitos, CA, USA: IEEE Computer Society, giu. 2018,

pp. 881–885. DOI: 10.1109/HPCC/SmartCity/DSS.2018.00146. URL: https://doi.ieeecomputersociety.org/10.1109/HPCC/SmartCity/DSS.2018.00146.

[23] Chen-Zhao Yang et al. "Preventing DeepFake Attacks on Speaker Authentication by Dynamic Lip Movement Analysis". In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 1841–1854. DOI: 10.1109/TIFS.2020.3045937.

[24] Jacob L Newman e Stephen J Cox. "Speaker independent visual-only language identification". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2010, pp. 5026–5029. DOI: 10.1109/ICASSP.2010.5495071.

[25] Jacob Newman e Stephen Cox. "Language Identification Using Visual Features". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20 (set. 2012), pp. 1936–1947. DOI: 10.1109/TASL.2012.2191956.

[26] Triantafyllos Afouras, Joon Son Chung e Andrew Zisserman. "Now You're Speaking My Language: Visual Language Identification". In: *Proc. Interspeech 2020*. 2020, pp. 2402–2406. DOI: 10.21437/Interspeech.2020-2921.

[27] Colah et al. "Understanding LSTM Networks". In: *colah.github.io/posts/2015-08-Understanding-LSTMs/* ().

[28] Emanuele Mezzi. "Visual Language Identification Con Modelli Neurali Ricorrenti". In: ().

[29] Malihe Bashirnezhad e Zargham Ghapanchi. "A Comparative Typology on Phonological System of Russian, English and Persian Languages". In: *International Journal of Educational Investigations*. 2017.

Ringraziamenti

In primo luogo vorrei ringraziare i miei relatori, il professore Abate e la Dottoressa Cascone, i quali mi hanno dato quest'opportunità di lavorare ad un progetto che mi ha molto affascinato e interessato dove mi hanno assistito e supportato per l'enorme complessità di argomenti che non avrei mai immaginato di affrontare da solo.

Ringrazio la mia famiglia che mi ha sempre sostenuto in qualsiasi momento in cui ne avevo bisogno per raggiungere i miei obiettivi, e questo è uno tra i più importanti. Non vi sarò mai grato abbastanza per tutti i sacrifici che avete dovuto affrontare per vedermi realizzare i miei sogni e di togliermi le preoccupazioni nell'affrontare qualsiasi cosa con tranquillità e serenità. Grazie veramente di tutto, vi voglio un mondo di bene.

Ringrazio a tutti coloro che ho conosciuto durante questo percorso universitario che mi hanno dato la spinta e il coraggio di affrontare qualsiasi cosa a testa alta credendo nelle mie capacità. Abbiamo condiviso momenti di gioia e sofferenza insieme vedendo sempre il lato positivo, spingendo sempre il massimo e andare anche oltre ai nostri limiti. Sono davvero felice di avervi conosciuti, di aver creato un'amicizia che può andare anche al di fuori dell'università e di aver percorso questa strada insieme dove spero di continuare a proseguirla anche dopo aver compiuto questo piccolo traguardo. Siete stati e continuerete ad essere degli ottimi compagni di avventura, vi voglio bene.

Ringrazio al mio gruppo di uscite e di *gaming* che mi hanno dato la giusta dose di svago nei tempi bui o di relax. Abbiamo condiviso momenti di divertimento e ignoranza, soprattutto imprecando su *League of Legends*, guardando video *meme*, film, serie tv ridendo a crepa pelle per ogni minima cosa che potesse far ridere. Vi ho condiviso il mio lato più scherzoso e ignorante che ci possa essere e sono felice di avere voi che ne facciate parte.

Infine, vorrei ringraziare i professori che mi hanno accompagnato durante gli anni delle superiori, che hanno riacceso in me la voglia di ricominciare a credere nelle cose che mi fanno sentire me stesso e che sono stati i primi a scovare le mie capacità.