

UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Scienze e Tecnologie

Corso di Laurea in Informatica (L-31)

ANALISI QUANTITATIVA E
PERCETTIVA DI VIDEO CREATI CON
GENERATORI AI

Relatore: Raffaella Lanzarotti

Correlatore: Alessandro D'Amelio

Tesi di:

Federico COSCIA

Matricola: 977772

Anno Accademico 2023-2024

a Celestina

Indice

Introduzione	1
1 Protocollo sperimentale	3
1.1 Stesura del protocollo	3
1.1.1 Obiettivo	3
1.1.2 Disegno sperimentale	4
1.1.3 Variabili indipendenti	5
1.1.4 Variabili dipendenti	7
1.1.5 Definizione dei questionari	8
1.1.6 Dati fisiologici, sguardo, espressioni facciali	12
1.1.7 Riepilogo	14
2 Generazione dei video fake	16
2.1 Funzionamento	16
2.1.1 Cos’è un Avatar	16
2.1.2 Diversi tipi di avatar	17
2.1.3 Pipeline di generazione	18
2.2 Valutazione delle soluzioni disponibili	18
2.2.1 DupDub	19
2.2.2 Synthesia.io	19
2.2.3 HeyGen	20
2.2.4 La scelta	21
2.3 Video generati da video scaricati dal web	22
2.3.1 Criterio di ricerca	22
2.3.2 Video trovati	22
2.3.3 Processing	23
2.3.4 Generazione dei video fake	24
2.3.5 Scelta degli avatar	25
2.3.6 Scelta delle voci	25
2.3.7 Inserimento del testo	27

2.3.8	Download dei risultati	28
2.4	Video generati con attori	29
2.4.1	Le limitazioni dell'approccio via web	29
2.4.2	Scrittura dei testi	30
2.4.3	Individuazione degli attori	30
2.4.4	Acquisizione dei video	31
2.4.5	Video processing	32
2.4.6	Audio processing	33
2.4.7	Generazione dei video fake	34
2.5	Pro e contro	36
2.5.1	Video generati da video scaricati dal web	36
2.5.2	Video generati con attori	37
2.5.3	La scelta	38
3	Raccolta dati	39
3.1	Sviluppo dell'interfaccia	39
3.1.1	Design di sviluppo	39
3.1.2	Selezione dei video	40
3.1.3	Fase di preparazione	42
3.1.4	Raccolta dei timestamp	43
3.1.5	Registrazione della webcam/dello schermo	43
3.1.6	Controllo dell'eye-tracker	45
3.2	Salvataggio dei dati raccolti	45
3.2.1	Dati demografici e questionari	45
3.2.2	Video	46
3.2.3	Eye-tracking	46
3.3	Dati fisiologici	46
3.3.1	Setup	46
3.3.2	Calibrazione	47
3.3.3	Acquisizione dei segnali	47
3.3.4	Salvataggio dei dati	47
3.3.5	Download dei dati	48
4	Analisi dei dati acquisiti	49
4.1	Preparazione dei dati	49
4.1.1	Suddivisione in gruppi	49
4.2	Espressioni facciali	50
4.2.1	Espressione media, minore, più estrema	50
4.2.2	Espressioni discrete	51
4.2.3	Creazione di un classificatore	54

4.2.4	Confronto delle AUs con UMAP	54
4.2.5	Riepilogo risultati	58
4.3	Dati Fisiologici	59
4.3.1	Picchi sistolici, Battito cardiaco	59
4.3.2	Calcolo dei BPM	60
4.3.3	Analisi del battito cardiaco	61
4.3.4	ElectroDermal Activity (EDA)	65
4.3.5	Analisi statistica	66
4.3.6	Analisi automatica	67
4.3.7	Riepilogo risultati	67
4.4	Eye-tracking	68
4.4.1	Fissazioni	69
4.4.2	Saccadi	71
4.4.3	Battiti di ciglia	77
4.4.4	Riepilogo risultati	78
4.5	Questionari	79
4.5.1	Risultati dei questionari	80
4.5.2	Questionari post-visione	81
4.5.3	Autovalutazione emotiva (PANAS)	84
4.5.4	Domande di comprensione	84
4.5.5	Tempi di risposta	87
4.5.6	Riepilogo risultati	88
5	Conclusioni	91
5.1	Lavori futuri	94
Bibliografia		96

Introduzione

Per chiunque abbia visitato qualunque sito di divulgazione o social media negli ultimi due anni, sarà stato impossibile non imbattersi in contenuti, fotografici o video, generati dall’Intelligenza Artificiale (IA). Tra queste tecnologie, si identificano sistemi in grado di clonare l’aspetto di una persona reale, permettendo di far dire a questa persona qualsiasi cosa, partendo semplicemente da un loro video o una loro fotografia, e un testo di riferimento. Tali video prendono il nome di video basati su “avatar”.

Più tecnicamente, un video basato su avatar è un video parlato in cui il movimento della bocca, della faccia, e talvolta anche del corpo, così come spesso anche il suono della voce, della persona rappresentata è stato generato tramite IA. Spesso gli avatar sono persone reali, le quali hanno messo a disposizione la loro figura affinché potesse essere animata. In altri casi, un avatar può anche essere creato a partire da immagini di persone non reali, generate a loro volta attraverso strumenti di IA generativa (es. Stable Diffusion, Midjourney, DALL-E) [3]. Si crea di fatto una copia digitale della persona raffigurata, con la possibilità di fargli dire qualsiasi cosa. Più precisamente, dato il testo del discorso e partendo da una fotografia o un breve video del soggetto, viene generato un video rappresentante il soggetto che espone ad alta voce il testo indicato.¹

Distinguiamo quindi due tipi diversi di video, i video reali, raffiguranti una persona reale registrati fisicamente, e i video “sintetici”, o “fake”, raffiguranti a loro volta una persona reale, ma la cui voce e il cui movimento del corpo sono stati realizzati tramite generatori IA. Questa tecnologia ha trovato grande fortuna nel mondo della pubblicità e della istruzione, dove sempre maggiori sono i costi per la registrazione di video in presa diretta. L’utilizzo di tali sistemi permette di realizzare video senza doverli registrare fisicamente, riducendo di molto i costi di produzione.

L’obiettivo di questo studio è valutare se questi video sintetici possono avere la stessa efficacia comunicativa di un video reale, oppure se la natura artificiale di tali video può risultare un ostacolo abbastanza grande nella comprensione e fruizione

¹Naturalmente la realizzazione di questi video richiede anche la presenza di una voce parlata. Per questo si utilizzano modelli generativi di tipo TextToSpeech, i quali a partire da un testo generano il suono di una voce. La voce generata può essere la voce stessa della persona raffigurata o anche una voce di servizio.

del contenuto video. Al fine di dare una risposta scientifica a questo quesito, è stato messo a punto un protocollo sperimentale che permette di valutare quantitativamente l'impatto della natura dei video sulla loro efficacia comunicativa, permettendo di dire se tale soluzione tecnologica, allo stato attuale, è adatta per essere adottata nel mondo dell'educazione multimediale.

Struttura dell'esperimento

L'esperimento è stato strutturato come segue:

1. Acquisizione di brevi video educativi reali (max. 5/10 minuti)
2. Generazione di doppioni fake a partire dai video reali acquisiti, dove il contenuto informativo è lo stesso, ma il video, così come la voce del soggetto, sono stati generati tramite IA
3. Presentazione di un esperimento fittizio, dove un campione di volontari visiona due video, valutando la loro esperienza di visione, ma, a insaputa dei partecipanti, uno dei due video mostrati è stato generato tramite IA
4. Sessione di domande di valutazione dell'esperienza su una scala da 1 a 5, e sessione di domande di comprensione a risposta multipla sui contenuti discussi nei video, per valutare il grado di comprensione dei contenuti proposti
5. Acquisizione di dati multimodali durante l'esperimento e in particolare durante la visione dei video, tra cui le espressioni del viso, il battito cardiaco, il livello di sudorazione, e il movimento degli occhi sullo schermo.
6. Analisi ed elaborazione dei dati acquisiti, alla ricerca di cluster indicativi di un fattore di rilevanza/non rilevanza della natura dei video per la comprensione e la fruizione del video visionato

Nel Capitolo 1 verrà spiegato nel dettaglio il protocollo sperimentale stabilito, al fine di rendere chiaro come è stato organizzato l'intero progetto. Nel Capitolo 2 verrà data una breve spiegazione semplificata di come questi tipi di video vengono generati, secondo la letteratura attuale, per poi entrare nel dettaglio della acquisizione dei video reali e la realizzazione dei video fake. Nel Capitolo 3 verrà approfondita la fase di sperimentazione e l'acquisizione dei dati. Nel Capitolo 4 verrà trattata l'elaborazione e lo studio dei dati raccolti. Infine, nel Capitolo 5, verranno tratte le dovute conclusioni in base a quanto trovato.

Capitolo 1

Protocollo sperimentale

Per cominciare, è bene presentare il protocollo sperimentale che è stato definito, in quanto questo è il filo conduttore che lega tutto il progetto di ricerca di tesi. Il progetto è stato svolto in collaborazione con il dipartimento di Psicologia dell'Università Cattolica del Sacro Cuore, con il professore Andrea Gaggioli, il quale ha supervisionato e contribuito alla stesura del protocollo sperimentale e il processo di acquisizione dei dati, in particolar modo per la definizione dei vari questionari utilizzati. Inoltre, ha messo a disposizione la strumentazione di eye-tracking, la struttura e lo spazio utilizzati per lo svolgimento dell'esperimento di seguito descritto, e un team di tesisti che hanno collaborato alla stesura del protocollo e l'individuazione dei partecipanti.

1.1 Stesura del protocollo

1.1.1 Obiettivo

L'obiettivo di questo studio è confrontare l'esperienza di apprendimento per due contenuti didattici distinti, presentati in versioni con docenti reali o sintetici e in varianti maschili o femminili, per esplorare l'impatto delle variabili di presentazione e genere del docente, e del contenuto stesso, sulla qualità percepita dell'apprendimento e sulle risposte emotive e fisiologiche. Per fare questo, è stato ideato un esperimento, dove un campione di partecipanti volontari prende visione di alcuni contenuti didattici, e tramite la somministrazione di questionari di valutazione dell'esperienza visiva, la somministrazione di domande di comprensione, e la misura di diversi dati fisiologici misurati durante la visione, come il battito cardiaco, il movimento degli occhi e la sudorazione, viene misurata e valutata l'esperienza di visione dei video. A insaputa dei partecipanti, solo alcuni dei video mostrati sono reali, la restante parte è stata generata tramite IA. In questo capitolo affronteremo nel dettaglio come l'esperimento è stato disegnato e realizzato per ottenere i risultati migliori.

1.1.2 Disegno sperimentale

Per la somministrazione dei video reali e sintetici ai soggetti volontari sono stati valutati due tipi di disegni sperimentali. Supponiamo di aver individuato un campione, ad esempio di 30 partecipanti, disposti a partecipare al nostro esperimento, e di avere una raccolta di video didattici da dover mostrare, divisi fra reali e sintetici. Vogliamo garantire che i video reali e i video sintetici siano visualizzati in modo equo tra tutti i partecipanti, ma come suddividere le acquisizioni? Si presentano subito due opzioni:

1. Ogni partecipante visiona un solo video: i partecipanti vengono divisi in due gruppi, un gruppo visionerà i video reali e l'altro i video sintetici. Si tratta del disegno sperimentale *between-subject*
2. Tutti i partecipanti visionano due video: ogni partecipante visiona sia un video reale che uno sintetico. Non vi è quindi una divisione in gruppi, si tratta del disegno sperimentale *within-subject*.

Between-subject

Nel disegno between-subject i partecipanti vengono suddivisi in due gruppi, e ogni partecipante prende visione di un solo video. Un partecipante vedrà un video reale se appartiene al gruppo dei video reali, mentre un partecipante vedrà un video sintetico se appartiene al gruppo dei video sintetici. Questo disegno sperimentale comporta delle acquisizioni molto veloci, in quanto ogni partecipante deve visionare un solo video, ma comporta alcuni dettagli tecnici.

Dimensione del campione Questo disegno sperimentale richiede un campione di dimensioni doppie rispetto al *within-subject*. Dividendo il campione in due gruppi, se si vuole garantire che ogni tipo di video venga visionato n volte, è necessario trovare un campione grande $2n$.

Generalizzabilità dei risultati Inoltre, questo disegno sperimentale richiede un campione di maggiori dimensioni per poter garantire la generalizzabilità dei risultati. Con questo disegno sperimentale per ogni partecipante valutiamo solo come questo reagisce a un contenuto reale, se appartiene al gruppo dei video reali, o a un contenuto sintetico, se questo appartiene al gruppo dei video sintetici. Per poter generalizzare i risultati, per cui, è necessario un campione di grandi dimensioni, così da poter eliminare le differenze tra i singoli individui, e poter trarre delle conclusioni generali.

Un vantaggio, però, di cui si deve tener conto di questo disegno sperimentale, è che visionando un solo video lo spettatore non può avere influenze da altri video durante la visione, producendo i dati più puliti ed accurati.

Within-subject

Nel disegno within-subject si elimina la suddivisione in gruppi, e a ogni partecipante sono mostrati due video, uno reale e uno sintetico. L'ordine di visione è bilanciato tra tutti i partecipanti, alternando quale viene visionato per primo, in modo da eliminare un possibile bias dovuto all'ordine di visione. L'ordine può anche essere stabilito in modo casuale, avendo un campione di grandi dimensioni che permette la distribuzione uniforme dei due casi. È possibile vedere al Paragrafo 1.1.3 un esempio di come questo ordine può essere stabilito in base al numero del partecipante. Secondo questo disegno, le acquisizioni richiedono più tempo da parte dei partecipanti, ma così facendo, si dimezza la dimensione del campione richiesta, in quanto ogni partecipante vede entrambi i tipi di video. Inoltre, questo disegno riduce il problema della generalizzabilità: Mostrando ad ogni partecipante sia un video reale che un video sintetico, è possibile misurare come la stessa persona reagisce ai due contenuti di natura diversa. Questo permette di ottenere risultati generalizzabili prima. Al contrario del between-subject, però, visionando due video la percezione del secondo video potrebbe essere influenzata dal primo video.

La scelta

In luce di queste osservazioni, è stato scelto il design *within-subject*. Ogni partecipante per cui visiona due video, di cui sempre uno reale e uno sintetico.

1.1.3 Variabili indipendenti

Stabilito il disegno sperimentale, sono state individuate le variabili dipendenti e indipendenti. Oltre alla natura del video (reale o sintetico), c'è da tenere conto anche del genere del docente, in quanto il genere può avere un influenza sull'esperienza di visione dei video [14], a seconda dello spettatore, per cui dobbiamo tenerne conto, e assicurarci che nel corso dell'esperimento non vi sia una predominanza di un genere sull'altro nei video mostrati. Inoltre, dal momento che abbiamo bisogno di mostrare due video, e non possiamo certamente mostrare lo stesso video due volte [1], abbiamo bisogno di due contenuti diversi che i docenti andranno a trattare nei video. Riasumendo, le variabili indipendenti in gioco che determinano il tipo di video che un partecipante può andare a vedere sono:

- Tipo di contenuto:
 - Contenuto A
 - Contenuto B
- Tipo di Presentatore:

Partecipante	Contenuto	Tipo	Genere	Contenuto	Tipo	Genere
	Video 1	Video 1	Video 1	Video 2	Video 2	Video 2
1	A	Sintetico	Maschile	B	Reale	Femminile
2	A	Reale	Maschile	B	Sintetico	Femminile
3	B	Sintetico	Maschile	A	Reale	Femminile
4	B	Reale	Maschile	A	Sintetico	Femminile
5	A	Sintetico	Femminile	B	Reale	Maschile
6	A	Reale	Femminile	B	Sintetico	Maschile
7	B	Sintetico	Femminile	A	Reale	Maschile
8	B	Reale	Femminile	A	Sintetico	Maschile
9	A	Sintetico	Maschile	B	Reale	Femminile
10	A	Reale	Maschile	B	Sintetico	Femminile
11	B	Sintetico	Maschile	A	Reale	Femminile
12	B	Reale	Maschile	A	Sintetico	Femminile
13	A	Sintetico	Femminile	B	Reale	Maschile
14	A	Reale	Femminile	B	Sintetico	Maschile
15	B	Sintetico	Femminile	A	Reale	Maschile
16	B	Reale	Femminile	A	Sintetico	Maschile

Tabella 1: Matrice sperimentale (ipotizzando un campione di 16 partecipanti).

- Reale (docente videoregistrato)
- Sintetico (personaggio generato da AI)
- Genere del Presentatore:
 - Maschile
 - Femminile

Ogni possibile combinazione di assegnamenti di queste tre variabili rappresenta un possibile video che un partecipante può visionare. Ad esempio, un assegnamento del tipo:

$$\text{Contenuto1} = A \wedge \text{Tipo1} = \text{Reale} \wedge \text{Genere1} = \text{Maschile}$$

rappresenta un video reale con un docente di genere maschile che espone il contenuto A. Sulla base di queste variabili andiamo a costruire una matrice sperimentale, ovvero una matrice che specifica, per ogni partecipante, il valore di ognuna di queste variabili per il primo e il secondo video che i partecipanti andranno a visionare durante l'esperimento, specificando di conseguenza i due video che ogni partecipante andrà a

visionare. Osserviamo che le tre variabili sono tutte e tre variabili binarie, ovvero possono assumere solo due valori. Questo ci permette di definire estremamente facilmente il vincolo che lega il primo e il secondo video: il secondo video sarà la negazione logica del primo, dove se $Contenuto1$, $Tipo1$, e $Genere1$ sono le variabili che specificano il contenuto, il tipo e il genere del primo video, per il secondo video le variabili associate assumeranno rispettivamente i valori $Contenuto2 = \overline{Contenuto1}$, $Tipo2 = \overline{Tipo1}$ e $Genere2 = \overline{Genere1}$. Per cui, ad esempio, sulla base dell’assegnamento indicato in precedenza, il secondo video sarebbe definito dai valori:

$$Contenuto2 = B \wedge Tipo2 = Sintetico \wedge Genere2 = Femminile.$$

Per costruire la matrice è sufficiente esplicitare tutte le possibili combinazioni di assegnamenti di queste tre variabili per il primo video, e per ogni combinazione ricavare i rispettivi valori per il secondo video, utilizzando il vincolo appena definito. Esplicitando tutte le possibili combinazioni garantiamo che, con un campione abbastanza grande, tutti i casi possibili sono coperti e non vi sono combinazioni predominanti che potrebbero sbilanciare i dati raccolti. Osserviamo che le variabili sono tre, per cui il numero di possibili combinazioni è di $2^3 = 8$, per cui dopo 8 partecipanti la matrice si ripeterà, si tratta quindi di una matrice periodica. È possibile vedere un esempio di questa matrice in Tabella 1, dato un campione di 16 partecipanti. È possibile notare la periodicità della tabella, dove i valori delle variabili per i partecipanti da 1 a 8 sono ripetuti per i partecipanti da 9 a 16. Si osservi che non è importante la combinazione di partenza, posto che il campione sia abbastanza grande da coprire tutte le possibili combinazioni. Per stabilire la dimensione del campione, infine, è sufficiente stabilire quante volte si desidera coprire tutte le combinazioni, e scegliere una dimensione appropriata. Ad esempio, se si vuole coprire tutti i casi possibili almeno cinque volte, servirà un campione di almeno $5 \cdot 8 = 32$ partecipanti.

1.1.4 Variabili dipendenti

In un esperimento, una variabile dipendente è una variabile che viene osservata e misurata, per studiare l’effetto che variazioni alle variabili indipendenti hanno su tale variabile. Nel nostro caso, si tratta di ciò che ci aspettiamo possa essere influenzato in uno spettatore dalla natura sintetica di un video. Le variabili dipendenti identificate sono:

- Esperienza emotiva durante la lezione (misurata con questionari psicometrici post-sessione)
- Qualità dell’apprendimento (misurata con test di comprensione post-sessione)
- Fissazioni e movimento dello sguardo (misurati con eye-tracker)

- Frequenza cardiaca e livello di sudorazione (misurati tramite sensore sul polso)
- Espressioni emotive facciali (monitorate tramite registrazione della webcam).

Le ragioni a supporto delle prime due variabili dipendenti sono che: vogliamo misurare come l'utilizzo di video sintetici piuttosto che reali possa alterare l'esperienza emotiva e la qualità dell'apprendimento in uno studente. Questi sono due aspetti fondamentali da studiare, se la comprensione o il coinvolgimento emotivo vengono meno utilizzando video artificiali questo vanifica il senso di utilizzarli. Per poter misurare queste due variabili sono stati definiti dei questionari di autovalutazione emotiva, di valutazione e di apprendimento, da somministrare durante l'esperimento.

1.1.5 Definizione dei questionari

I questionari sono stati definiti dal professore Andrea Gaggioli del dipartimento di Psicologia dell'Università Cattolica del Sacro Cuore. Alcuni dei questionari sono stati definiti per questo studio, altri sono questionari noti all'interno del mondo della Psicologia come il Positive And Negative Affect Schedule (PANAS) [9], per la autovalutazione del proprio stato emotivo, e il Video Engagement Scale (VES) [20], per la valutazione del livello di coinvolgimento durante la visione di un video.

Questionari pre-trattamento

Sono stati specificati dei questionari pre-trattamento, da somministrare ai partecipanti all'inizio dell'esperimento, prima della visione di qualsiasi video. Qui vengono raccolti semplici dati demografici come l'età, il genere, la nazionalità e il livello di istruzione di ogni partecipante. In seguito, è somministrato un questionario di autovalutazione del loro stato emotivo pre-trattamento, seguito da una domanda (in una scala da 1 a 5) su quanto frequentemente il partecipante fa utilizzo di contenuti multimediali per l'apprendimento. Più nello specifico, i questionari pre-trattamento consistono in, in questo ordine:

- Dati demografici (età, genere, nazionalità, livello di istruzione, occupazione)
- Questionario per valutare lo stato emotivo iniziale (Positive and Negative Affect Schedule - PANAS), che misura l'umore momentaneo
- Scala sulla frequenza di utilizzo di contenuti video per l'apprendimento (ad es., lezioni online, tutorial, corsi su piattaforme digitali)
- Scala per misurare l'uso della tecnologia in generale a scopi educativi (16 item)

Il questionario su scala per misurare l'uso della tecnologia consiste in:

1. Frequenza di utilizzo della tecnologia:

- *Quanto spesso utilizzi tecnologie digitali (ad esempio, app, piattaforme, dispositivi) per scopi educativi?*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)
- *Quanto spesso combini più tecnologie (ad esempio, video, quiz online, forum) nel tuo processo di apprendimento?*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)

2. Tipi di tecnologie utilizzate:

- *Utilizzo piattaforme di gestione dell'apprendimento (ad esempio, Moodle, Blackboard, Google Classroom).*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)
- *Utilizzo strumenti collaborativi (ad esempio, Microsoft Teams, Zoom, Google Docs) per scopi educativi.*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)
- *Utilizzo app o piattaforme educative specializzate (ad esempio, Coursera, Khan Academy, Duolingo).*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)

3. Scopi di utilizzo della tecnologia:

- *Utilizzo le tecnologie per accedere a materiali didattici (ad esempio, eBook, video, articoli).*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)
- *Utilizzo le tecnologie per collaborare con i miei pari o con gli insegnanti/datori di lavoro.*
(Mai, Raramente, A volte, Spesso, Molto frequentemente)

4. Impatto percepito sull'apprendimento:

- *Usare la tecnologia migliora la mia comprensione del materiale.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)
- *La tecnologia mi aiuta a rimanere organizzato e a gestire meglio il mio apprendimento.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

- *Apprendo meglio utilizzando le tecnologie rispetto ai metodi tradizionali.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

5. Adattabilità e competenza:

- *Mi adatto rapidamente alle nuove tecnologie introdotte a scopi educativi.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

6. Sfide e ostacoli:

- *A volte affronto difficoltà tecniche che ostacolano la mia esperienza di apprendimento.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)
- *L'accesso limitato a dispositivi o alla connessione internet influenza la mia capacità di usare la tecnologia per apprendere.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

7. Cointvolgimento e motivazione:

- *La tecnologia rende l'apprendimento più coinvolgente per me.* (Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)
- *Mi sento motivato a esplorare nuovi argomenti o competenze quando utilizzo la tecnologia.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

8. Integrazione nelle pratiche di apprendimento:

- *I miei insegnanti/datori di lavoro incoraggiano l'uso della tecnologia nel processo di apprendimento.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)
- *Integro regolarmente la tecnologia nella mia routine di studio.*
(Fortemente in disaccordo, In disaccordo, Né accordo né disaccordo, Accordo, Fortemente d'accordo)

Questionari post-video

Al termine della visione di ciascun video, viene poi somministrato un insieme di questionari. I questionari post-video consistono in un'autovalutazione del proprio stato emotivo in seguito alla visione, delle domande (in una scala da 1 a 5) sulla percezione della qualità dell'apprendimento e sulla valutazione del presentatore, un questionario sul livello di coinvolgimento durante la visione (Video Engagement Scale, o VES), e due domande sul livello di familiarità pregressa con l'argomento presentato e l'utilità percepita delle informazioni ricevute. Infine, è prevista una breve sessione di domande di comprensione a risposta multipla sui contenuti trattati nel video appena visionato, per misurare il grado di comprensione raggiunto. In un elenco puntato, i questionari post-video vengono presentati in questo ordine:

- Esperienza emotiva: Positive and Negative Affect Schedule (PANAS), per misurare i sentimenti positivi e negativi provati durante la video-lezione
- Percezione della qualità dell'apprendimento:
 1. *“Quanto hai trovato chiari i contenuti della lezione?”* (scala da 1 a 5)
 2. *“Quanto sono stati facili da comprendere i concetti presentati nella lezione?”* (scala da 1 a 5)
 3. *“Quanto hai trovato ben organizzata la presentazione dei contenuti?”* (scala da 1 a 5)
 4. *“Quanto pensi di aver appreso dai contenuti presentati?”* (scala da 1 a 5)
 5. *“Quanto ti senti sicuro/a di ricordare le informazioni apprese nella lezione?”* (scala da 1 a 5)
 6. *“Quanto ti senti preparato/a ad applicare i concetti appresi?”* (scala da 1 a 5)
 7. *“Quanto ritieni utile la lezione per il tuo apprendimento?”* (scala da 1 a 5)
 8. *“Quanto la lezione ha stimolato il tuo interesse per l'argomento?”* (scala da 1 a 5)
 9. *“Quanto ti senti motivato/a a saperne di più sull'argomento dopo aver visto la lezione?”* (scala da 1 a 5)
- Valutazione del presentatore
 1. *“Come descriveresti il modo di presentare del relatore?”* (scala da 1 a 5, dove 1 è “Molto impersonale” e 5 è “Molto coinvolgente”)

2. “Quanto hai trovato efficace il relatore nella trasmissione dei contenuti?”
(scala da 1 a 5)
 3. “Quanto ti è sembrato chiaro e sicuro il relatore durante la spiegazione?”
(scala da 1 a 5)
 4. “Quanto ti sei sentito/a in sintonia con il relatore?” (scala da 1 a 5)
 5. “Quanto il presentatore ti è sembrato naturale e realistico nella presentazione dei contenuti?” (scala da 1 a 5)
- Video Engagement Scale (VES)
 - Domande di familiarità e utilità
 1. “Quanto eri già familiare o a conoscenza dei contenuti mostrati nel video?”
(scala da 1 a 5)
 2. “Quanto ti è sembrato utile e/o informativo questo contenuto?”
(scala da 1 a 5)
 - Test di comprensione per valutare l'apprendimento oggettivo

I testi di comprensione per valutare l'apprendimento oggettivo consistono in un quiz da 5 domande a crocette, con 1 risposta giusta e 3 risposte sbagliate, sui contenuti trattati.

1.1.6 Dati fisiologici, sguardo, espressioni facciali

Le altre tre variabili dipendenti sono meno ovvie, e richiedono una breve sezione sul perché è stato valutato importante misurarle, e in seguito come sono state misurate.

Dati fisiologici

Il nostro corpo è in uno stato di costante reazione agli stimoli proposti. Oltre a quello che percepiamo, valori come il battito cardiaco, il ritmo respiratorio, il livello di sudorazione della pelle sono fortemente influenzati dalle nostre emozioni e dagli stimoli visivi (e in particolar modo anche sonori [18]) che ci sono presentati [11]. A partire da questi è quindi possibile studiare le risposte emotive associate agli stimoli presentati, nel nostro caso i video reali o sintetici. Tra questi, il livello di sudorazione della pelle, detto anche ElectroDermal Activity (EDA), è particolarmente utile, in quanto questo è fortemente collegato ai livelli di stress, tensione e reazione emotiva, specie durante l'apprendimento [10][19], e tende a essere fortemente reattivo.

Per la cattura dei dati fisiologici è stato predisposto un braccialetto Embrace Plus di Empatica Inc., indossato sul polso della mano non dominante di ogni partecipante.

Il braccialetto permette di misurare in maniera non intrusiva i segnali che ci interessano, quali battito cardiaco e livello di sudorazione della pelle. Per assicurarsi una lettura accurata dei dati richiesti, ai partecipanti viene richiesto di indossare il braccialetto per 10-15 minuti prima dell'inizio dell'esperimento, affinché la pelle del partecipante possa abituarsi al contatto con il sensore, come indicato dai manuali d'uso di Empatica. Il braccialetto è configurato e abilitato via applicazione via smartphone, e i dati sono salvati su un server Amazon AWS, da cui è possibile scaricare offline in qualsiasi momento tutti i dati raccolti. La gestione, acquisizione e salvataggio dei dati fisiologici è approfondito più nel dettaglio nella sezione 3.3.

Eye-tracker

I motivi per cui è di nostro interesse misurare e tracciare il movimento dello sguardo sono molteplici: innanzitutto, vista la presenza di video realizzati tramite IA, questo ci permette di identificare eventuali elementi di distrazione che potrebbero essere presenti in tali video, dovuti a possibili artefatti o imperfezioni nei video generati. Vista la natura ancora in via di sviluppo di questa tecnologia, un dettaglio sul viso del presentatore, il movimento o l'aspetto delle mani, per fare qualche esempio, potrebbero essere elementi di distrazione¹. Inoltre, la misura delle fissazioni e del movimento dello sguardo possono essere ulteriori indicazioni del grado di coinvolgimento emotivo e del grado di attenzione dei partecipanti verso il video proposto [15][16].

Per il tracciamento dello sguardo e la misura delle fissazioni è utilizzato l'eye-tracker EyeLink 1000, di SR Research Ltd. Questo eye-tracker permette di ottenere, con una frequenza di campionamento fino a 1000 campioni al secondo, la posizione dello sguardo sullo schermo ad ogni istante e il diametro della pupilla. Per l'utilizzo dell'eye-tracker è prevista una fase di calibrazione, svolta manualmente dallo sperimentatore insieme al partecipante prima di avviare l'esperimento. In seguito, l'eye-tracker è pilotato dal computer che guida il partecipante attraverso l'esperimento, per cui il tracciamento e il salvataggio dei dati raccolti è automatico. Per l'utilizzo dei dati di movimento dello sguardo, sarà necessario, inoltre, registrare lo schermo del computer dove sono presentati gli stimoli, così da poter associare un certo movimento dello sguardo, ad esempio, allo stimolo che lo ha causato.

Espressioni facciali

Come trovato da [21][22], le espressioni facciali, e le relative emozioni, sono un buon indicatore del livello di interesse e coinvolgimento di uno studente durante

¹Si osservi che non è necessariamente detto che queste distrazioni comportino un ostacolo nell'apprendimento, potrebbe evincersi dai risultati di questo studio che nonostante le fissazioni, dovessero essercene, i partecipanti non abbiano avuto problemi di comprensione. Questo arricchisce ancora di più l'importanza e l'utilità di questa variabile in questo studio.

l'apprendimento.

Per catturare le espressioni facciali, il partecipante viene registrato per tutta la durata dell'esperimento, previo consenso esplicito, tramite webcam, posta sullo schermo sul quale vengono visualizzati i video. In seguito, le espressioni facciali e le emozioni associate sono estratte dal video registrato utilizzando la libreria di Python `py-feat`.

1.1.7 Riepilogo

Per riepilogare, il disegno è di tipo *within-subject*. Ciascun partecipante visiona due video didattici:

- Uno recitato da un docente maschio e uno da una docente femmina.
- Uno in versione reale (docente videoregistrato) e uno in versione sintetica (generato dall'AI).

Le variabili indipendenti sono:

- Tipo di contenuto:
 - Contenuto A
 - Contenuto B
- Tipo di Presentatore:
 - Reale (docente videoregistrato)
 - Sintetico (personaggio generato da AI)
- Genere del Presentatore:
 - Maschile
 - Femminile

Mentre le variabili dipendenti sono:

- Esperienza emotiva durante la lezione (misurata con questionari psicometrici post-sessione).
- Qualità dell'apprendimento (misurata con test di comprensione post-sessione).
- Eye tracking (misure di fissazione e sguardo).
- Frequenza cardiaca e sudorazione (misurate con braccialetto munito di sensore).

- Espressioni emotive facciali (analizzate tramite registrazione della webcam ed estratte con la libreria Python `py-feat`).

Prima di cominciare, i partecipanti vengono accolti, e viene richiesta la firma del consenso informato sulla cattura di tutti dati sopra indicati. I partecipanti non sono informati della natura reale o sintetica dei personaggi nei video, per evitare bias sui dati raccolti. Prima di avviare l'esperimento, è prevista la preparazione delle misure fisiologiche:

- Battito cardiaco e sudorazione: Applicazione del braccialetto.
- Eye-tracking: Calibrazione delle misure oculari.
- Espressioni facciali: Posizionamento della webcam per il monitoraggio delle espressioni facciali.

Ogni partecipante visiona due video con contenuti diversi. Ciascun contenuto è presentato in una combinazione unica di caratteristiche: un contenuto sarà presentato da un docente maschile e l'altro da una docente femminile, uno sarà in versione reale e l'altro in versione sintetica. Il controbilanciamento copre anche l'ordine tra contenuto, genere e tipo di presentatore. Al termine, i partecipanti sono informati sui veri scopi dell'esperimento e sull'utilizzo dei dati raccolti.

Capitolo 2

Generazione dei video fake

In questo capitolo approfondiamo nel dettaglio il processo di generazione dei video fake realizzati per questo studio, partendo da una breve spiegazione semplificata su come queste tecnologie funzionano, seguendo poi con la selezione del servizio più adatto, per poi entrare nel dettaglio della creazione, confrontando due soluzioni diverse per la realizzazione dei video, dalla acquisizione dei video reali alla generazione dei video fake. Infine, le due soluzioni vengono confrontate e la scelta finale sui video da utilizzare viene fatta.

2.1 Funzionamento

In questa sezione andiamo ad esporre, in forma semplificata, il funzionamento di questi sistemi di generazione di video basati su avatar. Le tecniche e metodologie specifiche impiegate vanno oltre le competenze di una tesi triennale, inoltre la tecnologia specifica utilizzata varia da servizio a servizio, e, trattandosi di servizi commerciali, non è di dominio pubblico. Per questo motivo, andiamo ad esporre il funzionamento ad alto livello.

2.1.1 Cos'è un Avatar

Per “Avatar” si intende una copia sintetica dell’immagine di una persona. Questa copia permette la creazione di video fake, in cui la persona ritratta viene animata tramite Intelligenza Artificiale, generando, a seconda dei casi, i movimenti del volto, dello sguardo, della bocca, della testa, e anche del corpo. L’obiettivo di questa tecnologia è la generazione di un video parlato. Vi è quindi un testo di riferimento che viene utilizzato come guida per la generazione dei movimenti dell’avatar. Un avatar può essere realizzato a partire da una fotografia o un video raffigurante la persona, a seconda della tecnologia utilizzata. Naturalmente la potenza espressiva (così come

anche l'estensione di quanto è possibile animare) dipende dalla tecnologia utilizzata, e sono presenti sul mercato servizi diversi, in grado di offrire tecnologie con potenza diverse, a partire dai più semplici, i quali generano video a partire da una semplice fotografia, fino agli avatar più complessi, in grado di ricreare aspetto, voce e movimenti di una figura intera, così come il movimento dello sfondo, il tutto a partire solo da un breve video di essi.

2.1.2 Diversi tipi di avatar

Sul mercato si trovano diversi tipi di avatar, ognuno per uno scopo diverso.

Avatar Talking-Photo

Partiamo dal metodo più semplice per la generazione di avatar. Il modo più semplice con cui è possibile creare un avatar è a partire da una fotografia: utilizzando una semplice immagine raffigurante una persona, viene individuato il volto della persona ritratta, e tramite IA generativa vengono generati i movimenti dei muscoli del volto e della bocca, per dare l'impressione del parlato.

Questo è il metodo più semplice, ma tipicamente produce risultati molto poco realistici. Questo tipo di generazione è spesso detta “Talking-Photo”, appunto poiché crea l'impressione di una “fotografia parlante”: il resto dell'immagine rimane perfettamente statico, in particolar modo il resto del corpo della persona raffigurata, ovvero testa, capelli, braccia e corpo. I vantaggi di questo tipo di generazione sono che è molto semplice da utilizzare, e molto veloce, ma gli svantaggi sono la scarsa naturalezza e lo scarso realismo dei risultati prodotti.

Studio Avatar

La forma più diffusa di avatar sono gli “Studio Avatar”¹. Tipicamente ogni servizio di generazione di video basati su avatar fornisce un catalogo di avatar pronti all'uso di questo tipo. Uno Studio Avatar è un modello di una persona, in piano americano, o mezzo busto, privo di sfondo. La potenza espressiva di questi avatar varia da servizio a servizio, ma in genere questo tipo di avatar supporta movimenti più avanzati come il movimento dei capelli, della testa, del corpo, e talvolta anche delle braccia. Questi avatar possono essere realizzati a partire da un breve video della persona, dal quale viene rimosso lo sfondo. Per la generazione della voce è necessario utilizzare un modello di Text-To-Speech per la generazione del parlato. La traccia audio di voce parlata viene utilizzata come riferimento e guida per la generazione dei movimenti del corpo e della bocca.

¹Nessuno dei nomi usati per categorizzare gli avatar è “ufficiale”, ma diversi servizi tendono a utilizzare queste nomenclature per distinguere i vari tipi di avatar.

Avatar a sfondo reale

La forma più avanzata di avatar sono gli “Avatar a sfondo reale”. Questi sono avatar generati a partire da un breve video, raffigurante il soggetto inserito in un ambiente reale, in cui però lo sfondo è preservato, così come anche il suo eventuale movimento. In tale video, il soggetto è solitamente richiesto di parlare ad alta voce per un paio di minuti. Non sono importanti gli argomenti trattati, ma è importante scandire bene la pronuncia delle parole e l’uso di un’intonazione molto espressiva. L’avatar realizzato preserva lo sfondo presente nel video originale, ricreando, oltre che i movimenti della persona, quelli del volto, delle mani e del corpo, anche quello dello sfondo. Insieme al video, viene anche realizzato un modello Text-To-Speech della voce della persona raffigurata, permettendo di ricreare la voce della persona raffigurata, fornendo il livello di realismo più alto disponibile al momento. Naturalmente, questa potenza espressiva lo porta a essere il più dispendioso da utilizzare: per poter fare uso di questo tipo di avatar è necessario disporre di uno o più soggetti disposti a prestare la loro immagine, registrare i video di riferimento richiesti per la generazione degli avatar, con la qualità audio e video maggiore possibile, per poter garantire un risultato di alta qualità.

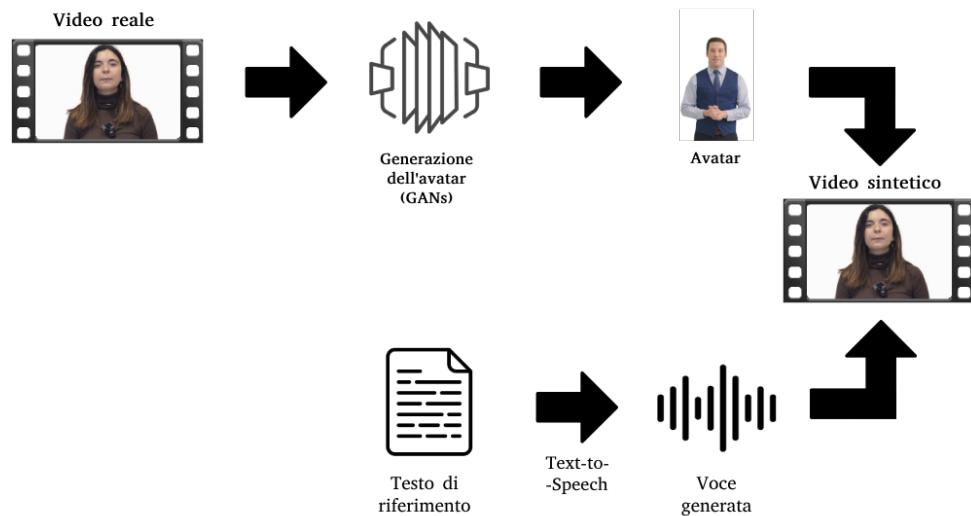
2.1.3 Pipeline di generazione

Presentiamo ora brevemente, in forma schematica, come avviene la generazione di un video fake basato su avatar, a partire da un testo e un video di riferimento reale. L’avatar viene realizzato a partire da una fotografia, o un breve video, del soggetto. Il processo di generazione fa tipicamente uso di modelli apprendimento basati su Generative Adversarial Networks (GANs). L’avatar generato può essere utilizzato per generare un video fake, a partire da un qualsiasi testo di partenza. Parallelamente, il testo di riferimento che si intende utilizzare per la generazione del video viene utilizzato per la generazione di una traccia vocale, tramite un modello di Text-to-Speech. Il file audio generato viene utilizzato come guida per l’avatar per la generazione del video, guidando i movimenti della bocca, del volto, e del corpo. È possibile vedere in Figura 1 uno schema di questo processo, nel caso di un modello generato a partire da un video di riferimento.

2.2 Valutazione delle soluzioni disponibili

Per la generazione dei video fake sono stati valutati tre applicativi diversi, forniti come Software-as-a-Service (SaaS):

- DupDub.com
- Synthesia.io



This cover has been designed using resources from Flaticon.com

Figura 1: Processo di generazione di un video fake, a partire da un video reale e un testo di riferimento.

- HeyGen.com

I criteri che sono stati valutati sono: la naturalezza dei movimenti generati, l'estensione dei movimenti generati, la possibilità di generare avatar personalizzati, la qualità del lip-sync², la qualità e la naturalezza della voce parlata generata, e il grado di realismo generale dei video generati. Vediamo per ordine i punti di forza e di debolezza identificati di ognuno, e come si è pervenuti alla scelta finale.

2.2.1 DupDub

DupDub si classifica come un prodotto “Talking-Photo”. DupDub trova i suoi punti di forza nell’essere molto semplice. Esso permette la creazione di un avatar personalizzato a partire da una sola fotografia di una persona. DupDub stato valutato come troppo impreciso per gli scopi di questa ricerca, limitandosi a generare solo i movimenti dei muscoli facciali e a malapena della testa, risultando non idoneo per questa ricerca.

2.2.2 Synthesia.io

Rispetto al precedente, Synthesia.io fornisce avatar in mezzo busto, ed è in grado di generare movimenti del viso, della testa, e anche del corpo, producendo risultati più

²Sincronizzazione tra il movimento delle labbra di un soggetto e il suono delle parole pronunciate.

naturali di DupDub. Gli avatar forniti sono stati generati a partire da persone reali, ed il servizio offre la possibilità di generare dei propri avatar personalizzati. I punti di debolezza individuati sono stati: la qualità del lip-sync e la qualità delle voci generate. In particolare, è risultato frequente il disallineamento tra il movimento delle labbra dell'avatar e il suono della voce generata. La voce inoltre è stata valutata come poco espressiva e poco naturale.

Nonostante questo, tale servizio poteva essere un buon candidato per la ricerca, ma è stato scartato a causa del piano offerto, in quanto offriva un servizio ad abbonamento basato su minuti di video generati.

2.2.3 HeyGen

Sin dal primo sguardo, HeyGen si è dimostrato essere al di sopra di tutti gli altri. HeyGen si distingue dalle sue controparti supportando video con sfondi reali, e generando movimenti del corpo avanzati come il movimento delle braccia e il gesticolamento delle mani.

Per la generazione dei video, HeyGen offre due soluzioni:

1. Generazione con i modelli di avatar e di voce forniti dalla piattaforma
2. Creazione di un avatar personalizzato, che a partire da un video di riferimento, clona l'aspetto e la voce della persona raffigurata, mantenendo lo sfondo raffigurato

Selezionato il metodo di generazione, si inserisce un testo di riferimento, a partire dal quale verrà generato il video fake.

Generazione con i modelli di avatar e di voce forniti dalla piattaforma

La piattaforma fornisce un catalogo di avatar e di modelli di voce già pronti per l'utilizzo. Con questi è possibile generare un video fake utilizzando soltanto un testo di riferimento. È il metodo più veloce per la generazione di video fake, poiché non ha bisogno di un video di riferimento per la creazione di un avatar ad-hoc, e permette di iniziare a generare video immediatamente. Si è, d'altra parte, limitati dall'aspetto degli avatar forniti dalla piattaforma. Gli avatar forniti sono privi di sfondo, per cui il video generato presenta l'avatar al centro dell'inquadratura, posto su uno sfondo bianco. Per la generazione è necessario identificare l'avatar che si intende utilizzare, ed identificare il modello di voce più adatto all'avatar scelto, tra quelli forniti dalla piattaforma. La piattaforma fornisce modelli di voce compatibili con tutte le lingue del mondo, ma tra tutti i modelli forniti, i modelli in lingua inglese sono i più naturali.

Generazione con avatar personalizzato

È possibile creare un avatar personalizzato a partire da un video di riferimento. A partire da tale video, la piattaforma identifica la persona raffigurata nel video, e ne crea un suo avatar. L'avatar creato non è privo di sfondo, bensì è inserito nello stesso sfondo in cui è stato registrato il video originale, aumentando il grado di realismo del video prodotto. Nella creazione di questo avatar è anche clonata la voce del soggetto rappresentato, per cui per la generazione dei video fake verrà utilizzata la voce della persona raffigurata, eliminando il problema di dover scegliere il modello di voce più adatto. Inoltre, la piattaforma si è dimostrata in grado di apprendere bene l'inflessione e l'accento della persona raffigurata, producendo risultati naturali indipendentemente dalla lingua parlata. Con questa soluzione è per cui possibile usare anche video in lingua italiana senza compromettere la qualità del risultato prodotto. C'è solo un dettaglio di cui tener conto, per la creazione di un avatar personalizzato è necessario il consenso esplicito in formato video della persona raffigurata che acconsente verbalmente l'utilizzo della sua immagine per la creazione di un avatar sulla piattaforma.

2.2.4 La scelta

Per queste ragioni, tra le opzioni valutate, HeyGen è stato valutato come il più adatto, in termini di qualità e naturalezza dei risultati prodotti, ed è stato quindi scelto come soluzione per questa ricerca. Sono state valutate entrambe le opzioni offerte da HeyGen per la generazione dei video fake, i cui approcci vengono approfonditi nella prossima sezione. Un altro fattore che sicuramente ha giocato a suo favore è stato anche il piano offerto, il quale permette di generare infiniti video durante il periodo di abbonamento, posto che questi siano sotto i cinque minuti di durata.

Profilo dei video fake

Il video che siamo interessati a generare ha le seguenti caratteristiche: un video raffigurante una persona che parla, inquadrata a mezzo busto, privi di movimenti di macchina o cambi di inquadrature, e privi di animazioni o scritte che compaiono a corredo. Il video può essere a sfondo bianco (generazione con avatar forniti dalla piattaforma) o con uno sfondo reale (generazione con avatar personalizzato).

Viste le due possibili soluzioni per la generazione dei video fake, sono state valutate due soluzioni diverse per l'acquisizione dei video reali di riferimento:

- Generazione di video fake a partire da video scaricati dal web, facendo utilizzo degli avatar già forniti dalla piattaforma, così facendo si ottengono soggetti diversi tra i due tipi di video

- Generazione di video fake a partire da video registrati con attori, realizzando avatar personalizzati così da avere lo stesso soggetto tra i due tipi di video

Vediamo ora i dettagli di entrambe le soluzioni, riportando l'approccio seguito, e valutando infine i pro e i contro di ogni soluzione.

2.3 Video generati da video scaricati dal web

2.3.1 Criterio di ricerca

È stata utilizzata per la ricerca dei video la piattaforma YouTube. Il criterio di ricerca usato è stato: cercare i video più simili possibili ai video fake che siamo in grado di generare, così da minimizzare le differenze tra video reali e video fake. Minimizzando le differenze tra video reali e video fake massimizziamo le possibilità che diverse percezioni dei video visualizzati siano dovute solo alla natura (reale o fittizia) dei video visualizzati e non ad altri dettagli come ambientazione, soggetto, etc. Per tali ragioni, sono stati cercati video:

- frontali, con un soggetto al centro su sfondo bianco
- con nessun movimento di macchina o cambi di inquadrature
- autodescrittivi, in altre parole non vengono utilizzate immagini, slide o grafici di supporto che vengono esplicitamente referenziati dallo speaker³
- con il minor numero di scritte o immagini che compaiono a corredo, preferibilmente nessuna
- con i sottotitoli preferibilmente inseriti a mano dall'autore del video, in modo da poter scaricare il copione associato al video più facilmente, per la generazione del video associato
- preferibilmente di durata inferiore ai cinque minuti

2.3.2 Video trovati

Durante il periodo di ricerca, sono stati trovati quattro video che soddisfano i criteri stabiliti:

- “How to make a GREAT impression - Presentation Tips” di Expert Academy (<https://youtu.be/lZg6H0WqPVY>)

³Questo perché contenuti esplicitamente referenziati dallo speaker reale non sarebbero presenti nel corrispettivo video fake, creando un'incongruenza e rendendo il video fake inefficace.

- “*How to start a pitch or presentation*” di Dominic Colenso (<https://youtu.be/P2LwuF7zn9c>)
- “*How to start a presentation*” di Expert Academy (<https://youtu.be/Lrj1W00kkws>)
- “*How to Get Over Your Fear of Public Speaking*” di Expert Academy (<https://youtu.be/So3Z93hEPDk>)

I video sono stati scaricati utilizzando il tool open source `yt-dlp`⁴.

2.3.3 Processing

I video individuati non corrispondevano tutti perfettamente alle specifiche richieste, per cui per poterli integrare nella ricerca è stato necessario fare del pre-processing.

Trimming

Tutti i video individuati presentavano un introduzione e una coda al video, con musiche, scritte o elementi animati. I video individuati sono per cui stati tagliati, in modo da eliminare gli elementi non utili al nostro studio, e mantenere solamente la parte di video parlata. Per il trimming dei video è stato utilizzato il tool open source gratuito `ffmpeg`⁵, così da favorire un’elaborazione veloce e priva di operazioni di re-encoding ove possibile.

Pulizia dello sfondo

Alcuni dei video individuati presentavano alcuni elementi grafici a comparsa durante la parte parlata del video, come grafici o piccole scritte. Questo è stato valutato come accettabile visto che tali elementi non venivano referenziati esplicitamente dal speaker, e comparivano solo in sovrapposizione dello sfondo.⁶ Questo ha permesso la rimozione di tali elementi aggiuntivi tramite una semplice operazione di video-editing, detta mascheramento.

Mascheramento Si identifica un fotogramma dell’immagine dove non vi sono elementi a coprire la parte dello sfondo interessata, e si salva tale fotogramma come file a parte. Questo fotogramma “pulito” è detto *clean plate*. Dal momento che lo sfondo

⁴<https://github.com/yt-dlp/yt-dlp>

⁵<https://www.ffmpeg.org>

⁶Ricordiamo che tutti i video individuati presentano uno sfondo bianco uniforme, che non cambia nel tempo.

è statico, ovvero non cambia nel tempo, il clean plate funge da copia pulita dell'immagine, che possiamo utilizzare per coprire qualunque elemento in sovrapposizione dello sfondo. Con un qualunque programma di editing, si sovrappone il clean plate alla porzione temporale di video in cui compare l'elemento da rimuovere, ad esempio una scritta, e si effettua poi una maschera, che va a ritagliare il clean plate. Come una toppa, il clean plate mascherato copre il testo in sovra-impressione, rimuovendolo dal video. È possibile vedere un esempio di questa operazione in Figura 2.



Figura 2: Una operazione di mascheramento con clean plate in Adobe Premiere Pro.

Estrazione del testo

Per poter generare i doppioni fake, è stato estratto il testo associato al parlato presente nei video individuati. È stato utilizzato il sito web gratuito <https://downsub.com> per scaricare i sottotitoli già forniti da YouTube. La maggior parte dei video presentavano dei sottotitoli ufficiali, ovvero inseriti direttamente dagli autori dei video. Per gli altri, sono stati scaricati i sottotitoli generati automaticamente da YouTube, utilizzando quindi di fatto il motore SpeechToText integrato di YouTube.

In ogni caso, tutti i sottotitoli scaricati sono stati poi revisionati a mano per eliminare refusi, errori di battitura o di trascrizione, e per eliminare elementi non parlati o associati alle parti di video che sono state tagliate via. Questi file di sottotitolo sono tutto il necessario per generare i video fake.

2.3.4 Generazione dei video fake

Dal momento che non è possibile ottenere il consenso esplicito dei soggetti rappresentati per la generazione di un avatar personalizzato, è necessario ricorrere agli avatar già forniti dalla piattaforma HeyGen per la generazione dei video fake. Tale processo prevede la selezione di un avatar tra quelli forniti dalla piattaforma, la selezione di una voce tra i modelli Text-To-Speech disponibili per generare il parlato, e infine l'inserimento del testo di riferimento.

Per ognuno dei video real individuati sono stati generati due video fake, uno con un avatar di genere maschile e uno con un avatar di genere femminile. Per la generazione di un video fake è stata seguita la seguente procedura, per ogni video real:

1. Scelta di un avatar
2. Scelta del modello di voce più adatto all'avatar scelto
3. Se non è stata trovata una coppia avatar-voce convincente tornare al passo 1 passando al prossimo avatar
4. Inserimento del testo estratto dal video real
5. Fine-tuning del testo per migliorare intonazione, pronuncia e pause
6. Revisione del risultato, ripetere il passo 5 se necessario
7. Ripetizione del processo con un avatar del genere opposto

2.3.5 Scelta degli avatar

Il punto di partenza per la generazione di un video fake è la scelta dell'avatar da utilizzare, ovvero la persona che verrà animata per realizzare il video parlato. La piattaforma HeyGen mette a disposizione una sua selezione di avatar proprietari, disponibili a tutti gli utenti del servizio, per realizzare i video fake. Gli avatar sono figure di persone a mezzo busto o in primo piano, prive di sfondo. È possibile vedere in Figura 3 un esempio ridotto della schermata di selezione degli avatar forniti da HeyGen. Tra gli avatar sono disponibili look molto variegati, tra cui figure in abiti formali, completi, in camice, abiti da lavoro, abiti casual, etc. Per il nostro studio, sono stati considerati avatar con un look semi-formale o casual.

La piattaforma offre anche la possibilità di realizzare un proprio avatar, a propria immagine e somiglianza, ma non è stato possibile nel nostro studio usufruire di questa feature, avendo utilizzato come video real video di terzi.⁷

2.3.6 Scelta delle voci

Come anticipato, la scelta degli avatar non è stata fatta in modo indipendente, ma è stata fatta in funzione dei modelli di voce forniti dalla piattaforma HeyGen. Difatti, anche se il video generato è visivamente impeccabile, una voce innaturale o non calzante all'avatar selezionato è in grado di rompere completamente l'illusione,

⁷È richiesto il consenso esplicito del soggetto rappresentato per realizzare un avatar a sua immagine.

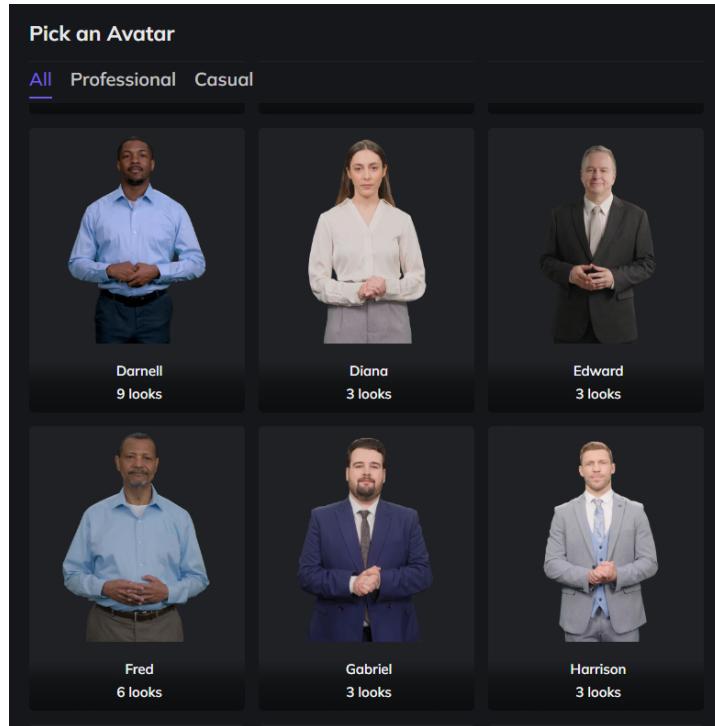


Figura 3: Esempio di schermata di selezione dell’avatar della piattaforma HeyGen.

risultando chiaramente artificiale, o può essere un fattore di distrazione, in grado di impedire la fruizione normale del contenuto. Riconosciamo come il giudizio di una proprietà come una voce “calzante al soggetto identificato” può essere fortemente personale, così come anche fortemente umana, e meriterebbe uno studio approfondito a parte. Per i nostri scopi, la scelta è stata guidata dal giudizio umano.

Filtraggio tramite categorie di voci

La piattaforma mette a disposizione un catalogo di voci molto variegato, suddiviso per categorie. Le categorie fornite sono visibili in Figura 4, e sono: genere (Maschio, Femmina), età (Child, Young adult, Middle-aged, Old), e “use case” (Conversazionale, Pubblicità e social, Informativo ed educativo, Narrativo). Sono state innanzitutto filtrate le voci selezionando il genere appropriato e la fascia di età appropriata per l’avatar selezionato. In aggiunta, sono state favoreggiate voci categorizzate come a scopo “Informativo ed educativo”, ma se necessario sono state valutate anche voci con altri use-case.

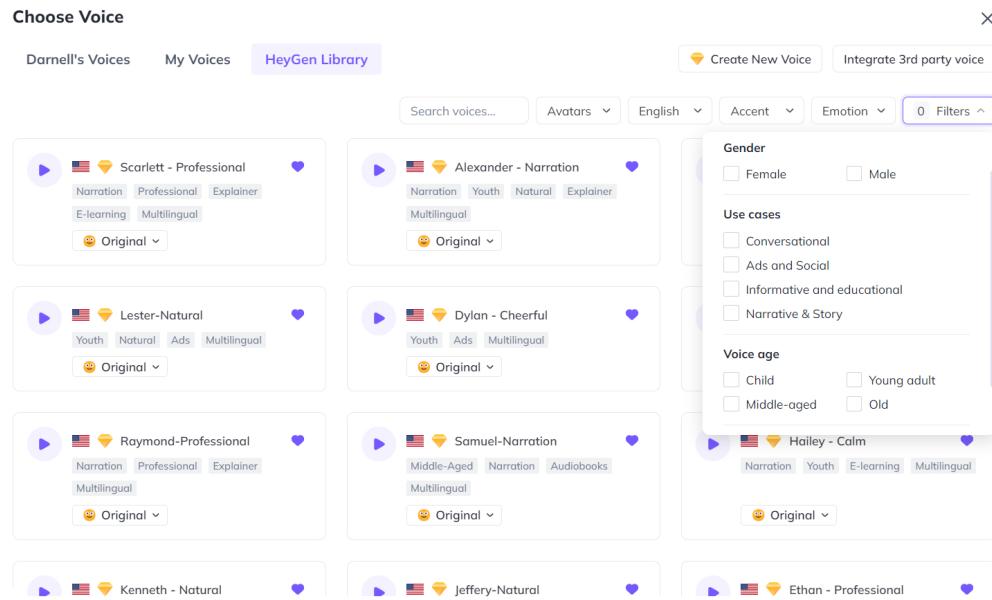


Figura 4: Schermata di selezione della voce sulla piattaforma HeyGen.

Processo di selezione

Isolate le possibili voci candidate, è stato generato un video per ogni voce. È stata selezionata poi, tra le candidate, la voce che, visionando il video generato, al giudizio umano è parsa più naturale e convincente con l'avatar selezionato. Se nessuna voce delle voci provate tra quelle fornite dalla piattaforma HeyGen è risultata convincente, l'avatar è stato scartato.⁸

Lingua

Tutti i video sono stati generati in lingua inglese poiché, sebbene HeyGen fornisca modelli di voci italiane, questi al tempo della ricerca erano limitati in numero e di qualità fortemente limitata rispetto alle controparti anglosassoni.

2.3.7 Inserimento del testo

L'ultimo passaggio per generare un video fake è l'inserimento del testo da far esporre all'avatar. Nel nostro caso, si tratta del testo estratto dai video real, come spiegato in

⁸C'è da notare come con il tempo la piattaforma si è evoluta, e al tempo della scrittura di questo documento, HeyGen fornisce insieme agli avatar una pre-selezione di voci adatte all'avatar selezionato. Questo processo risulterebbe per cui molto semplificato.

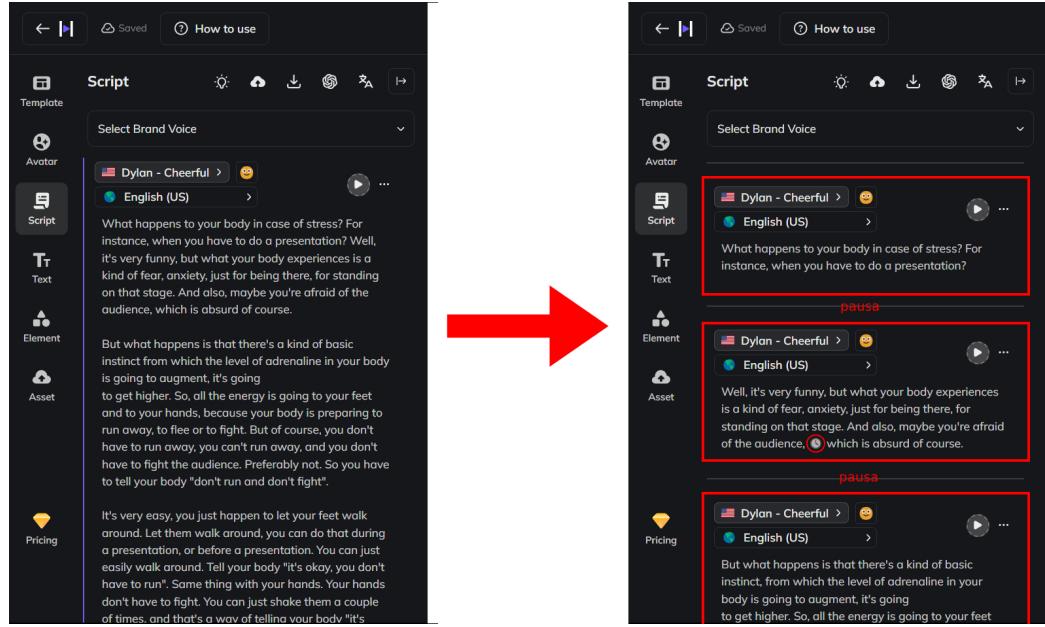


Figura 5: Separazione del testo in paragrafi per introdurre pause naturali.

2.3.3. I punti chiave di questo passaggio sono: l'introduzione di pause per un flusso naturale del discorso, e la specifica di particolari pronunce, ove necessario.

Introduzione di pause

La piattaforma HeyGen permette di introdurre pause nel discorso in modo naturale separando il testo in “paragrafi”. I paragrafi sono blocchi indipendenti di testo a partire dai quali è generata la voce. Tra un paragrafo e l’altro viene inserita automaticamente una piccola pausa, permettendo un flusso naturale del discorso. È possibile vedere un esempio di separazione in paragrafi in Figura 5.

Revisione del risultato

A partire dal testo inserito viene generato l’audio della voce, che farà da guida per la generazione dei movimenti dell’avatar, come spiegato in 2.1. Prima di avviare la generazione del video è possibile generare un’anteprima della voce. Se non si identificano problemi di pausa o di pronuncia, si fa partire la generazione del video.

2.3.8 Download dei risultati

Una volta generati i video, questi sono visualizzabili sulla piattaforma e scaricabili gratuitamente in formato 720p o 1080p. I video sono stati scaricati in formato 1080p.

2.4 Video generati con attori

2.4.1 Le limitazioni dell'approccio via web

L'approccio appena presentato è molto semplice e conveniente, in quanto non richiede la registrazione di video appositi per poter realizzare dei video fake, ma presenta delle forti limitazioni. Innanzitutto, il contenuto.

Complessità del contenuto

A livello contenutistico, si è limitati dai video che si è in grado di trovare in rete che soddisfano i requisiti richiesti (sfondo bianco, nessun cambio di inquadratura, ecc.). Nonostante la vastità della piattaforma YouTube, i criteri richiesti sono molto specifici, per cui i video trovati sono stati valutati molto limitati, ed in particolare il contenuto, seppur educativo, è stato valutato come di semplice comprensione. Il protocollo sperimentale stabilito prevede la somministrazione di alcune domande di comprensione sui contenuti affrontati, ma se questi sono troppo semplici, la somministrazione di tali domande diventa inefficace nel valutare l'efficacia dell'esperienza. I contenuti esposti nei video devono avere il giusto livello di complessità per essere comprensibili dallo spettatore, ma stimolarne l'attenzione, mettendo alla prova le sue capacità di comprensione.

Differenza di soggetti tra video reali e video fake

Un'altra limitazione, più tecnica, è la presenza di soggetti diversi tra un video reale e il video fake associato. Come già detto, individuare i video sul web riduce i tempi di acquisizione dei video, ma ci costringe a usare gli avatar forniti dalla piattaforma HeyGen per la generazione dei video fake. Non è possibile creare degli avatar personalizzati (quindi a immagine e somiglianza) dei video reali, poiché sarebbe necessario il consenso esplicito formato video dei soggetti raffigurati, di cui ovviamente non si può avere a disposizione, essendo i video stati trovati in rete. Bisogna, per cui, utilizzare gli avatar forniti dalla piattaforma HeyGen per realizzare i video. I video reali e i video fake presentano così soggetti diversi.

Seppure entrambi i video sono privi di sfondo, non hanno cambi di inquadratura, immagini o testi di corredo, ecc., non è esattamente vero che l'unica differenza tra i due tipi di video è la loro natura (reale o fittizio). Tra due video che presentano lo stesso argomento con le stesse identiche parole, se i due presentatori sono diversi vi può essere una differenza di percezione nello spettatore, in termini di livello di attenzione, interesse verso l'argomento, e fattore di comprensione degli argomenti esposti. Questo introdurrebbe una variabile esterno che potrebbe influenzare e sporcare i risultati

ottenuti, creando differenze tra i dati non dipendenti dal fenomeno che si intende studiare.

La soluzione

Entrambi questi problemi trovano una soluzione comune: registrare dei video ad-hoc con attori. Così facendo, è possibile avere il controllo sul livello di qualità, complessità e sull'argomento trattato nei contenuti esposti, scrivendo un copione preciso da seguire. Inoltre, è possibile fare uso degli avatar personalizzati, garantendo di avere lo stesso soggetto tra video real e video fake, poiché possiamo ottenere il consenso esplicito per la creazione degli avatar di chi si presta come soggetto per il video.

2.4.2 Scrittura dei testi

La realizzazione dei video con attori parte dalla scrittura dei testi che verranno esposti. Come previsto dal protocollo sperimentale, sono stati realizzati due testi, inequenti a due argomenti diversi, di media lunghezza. Questo progetto è stato svolto in collaborazione con l'Università Cattolica del Sacro Cuore, e i testi necessari per la realizzazione dei video sono stati scritti da un tesista dell'Università Cattolica, Matteo Scarinzi, come parte del suo progetto di tesi, seguito dal Prof. Andrea Gaggioli, docente della facoltà di Psicologia della Cattolica.

2.4.3 Individuazione degli attori

Una volta scritti i testi, è stato necessario individuare gli attori disposti a prestare il loro aspetto e la loro voce per registrare i video reali, a partire dai quali vengono realizzati i video fake.

Sono stati cercati un uomo e una donna, di età simile, preferibilmente con già qualche esperienza nell'esporre, spiegare, e parlare in pubblico, per facilitare le operazioni di ripresa dei video. È stato valutato come importante che i due attori avessero età simile per evitare che una chiara differenza di età potesse avere un'influenza diversa sull'esperienza di visione dei video, in termini di comprensione e grado di attenzione e coinvolgimento.

La ricerca è stata effettuata dal tesista della Cattolica Matteo Scarinzi all'interno del corpo docenti e dottorandi della facoltà di Psicologia della Cattolica, seguendo l'intuizione che docenti e dottorandi abbiano buona esperienza nell'esporre argomenti ad alta voce e parlare al pubblico. Sono stati individuati come attori:

- Michele Paleologo, per la parte maschile, dottorando della facoltà di Psicologia dell'Università Cattolica del Sacro Cuore

- Marta Pizzolante, per la parte femminile, dottoranda della facoltà di Psicologia dell’Università Cattolica del Sacro Cuore

2.4.4 Acquisizione dei video

I video sono stati registrati presso la sede dell’Università Cattolica del Sacro Cuore in Via Buonarroti, 30 a Milano, nell’aula privata del Prof. Andrea Gaggioli, docente di Psicologia della Cattolica e collaboratore di questo progetto.

I video sono stati registrati con luce artificiale, così da avere il pieno controllo della luce in scena, e non dipendere dalle condizioni meteo, come il movimento delle nuvole e del sole per la quantità di luce presente in scena durante i video e tra un video e l’altro. Per compensare la luminosità ridotta delle luci della stanza è stata utilizzata una ring light⁹, posta davanti la telecamera, così da illuminare gli attori.

Per la acquisizione dei video è di fondamentale importanza che i testi scritti siano seguiti parola per parola dagli attori, per evitare che differenze tra diverse modalità di esposizione possano influenzare i risultati ottenuti. Per garantire ciò senza che gli attori debbano imparare a memoria i testi scritti agli attori, è stato acquistato e utilizzato un teleprompter¹⁰. Per la registrazione dell’audio è stato utilizzato un microfono lavalier wireless. Il setup utilizzato per la registrazione è visibile in Figura 7.

Nonostante la capacità del servizio di generazione dei video fake di replicare lo sfondo nei video catturati, i video sono stati registrati di fronte a uno sfondo bianco, per evitare la presenza di possibili elementi di distrazione nello sfondo. È possibile vedere in Figura 6 un esempio della sessione di registrazione.

Video di riferimento per la creazione degli avatar

Per la realizzazione dei video fake è stato innanzitutto realizzato un primo video richiesto, richiesto dalla piattaforma HeyGen, per la creazione dell’avatar personalizzato (Instant Avatar). La piattaforma richiede un video di breve-media durata, 2-3 minuti, dove si può parlare di qualsiasi argomento. Il video è utilizzato come punto di partenza per creare l’avatar, estraendo da tale video l’aspetto del soggetto, dello sfondo, così come i movimenti del soggetto e il suono della sua voce. Tale video è stato registrato per ogni attore nello stesso identico setup, così che l’avatar realizzato fosse visivamente identico ai corrispettivi video reali. In tale video vengono inserite frequenti pause tra una frase e l’altra, così da garantire un risultato naturale nella creazione dell’avatar, come richiesto dalla piattaforma. Questo porta a un flusso del

⁹Lampada a forma di anello a luce bianca, molto luminosa, che permette attraverso la sua forma caratteristica di illuminare uniformemente il viso senza creare ombre dure sul volto.

¹⁰Un meccanismo a specchio che permette, a chi si pone davanti alla telecamera, di leggere il testo senza distogliere lo sguardo dalla camera.



Figura 6: Sessioni di registrazione dei video reali con attori, con: (a) Michele Paleologo nella parte maschile, (b) Marta Pizzolante nella parte femminile.

discorso molto innaturale, non rendendo possibile adoperare i video reali per la realizzazione dell'avatar, richiedendo la registrazione di un video dedicato a parte. Tale video è stato processato allo stesso modo di tutti gli altri, così che fosse visivamente e qualitativamente uguale ai corrispettivi video reali.

2.4.5 Video processing

I video registrati sono stati elaborati per migliorare la qualità dell'immagine. Sono stati applicati un filtro di riduzione del rumore e un aumento della luminosità dell'immagine. Per l'elaborazione dei video è stato utilizzato il programma gratuito Da Vinci Resolve.

Riduzione del rumore

I video sono stati registrati in condizioni di luminosità ristretta, utilizzando la luce artificiale dell'aula a disposizione e una ring light come luce di supporto. Per questo motivo, l'immagine presentava del leggero rumore, che è stato rimosso tramite un filtro di riduzione del rumore, incluso nel programma utilizzato. Il filtro di riduzione del rumore è stato controbilanciato con un filtro di sharpening¹¹, per evitare la perdita di dettagli dovuta alla riduzione del rumore. È possibile vedere un esempio del risultato di tale processo in Figura 8¹².

¹¹Un aumento della nitidezza dell'immagine, per rendere i dettagli più evidenti.

¹²Si tratta di un esempio molto sottile, e vista la dovuta compressione dell'immagine, necessaria per l'inserimento delle figure in questo documento, e la sua successiva stampa, potrebbe non essere facilmente apprezzabile la differenza tra le due immagini.



Figura 7: Setup di registrazione dei video fake, con videocamera, teleprompter e ring light.

Color Grading

L'immagine è stata migliorata con una correzione del bilanciamento del bianco, per correggere i colori, un leggero aumento della saturazione e della nitidezza, per correggere il profilo neutro della fotocamera utilizzata per le riprese, ed è stata aumentata la luminosità dell'immagine. È possibile vedere un esempio risultato finale in Figura 9.

I video sono stati esportati in formato .mp4, con codifica H.264 con un target di bitrate di 2048 kb/s, a 23.976 fps.

2.4.6 Audio processing

L'audio registrato è stato normalizzato a un valore di loudness di -23 LUFS (± 0.5 LU) integrati (I), secondo lo standard Europeo EBU R128 di distribuzione audio in ambito broadcast. Inoltre, l'audio è stato pulito con un filtro di riduzione del rumore di fondo, ed è stata applicata una curva di equalizzazione (Figura 10) e una leggera compressione per rendere l'ascolto più stabile, naturale e gradevole.

L'audio è stato convertito in formato AAC mono a 128 kb/s per poter essere inserito nel contenitore .mp4.



Figura 8: Esempio di filtro di riduzione del rumore dell’immagine nei video con attori.



Figura 9: Esempio della correzione dell’immagine effettuata su video con attori.

2.4.7 Generazione dei video fake

La generazione dei video fake risulta molto semplificata nel caso di video realizzati con attori. Una volta ottenuto il consenso esplicito degli attori che si sono prestati per la realizzazione dei video, tramite la piattaforma HeyGen è possibile effettuare la creazione degli avatar personalizzati (Instant Avatar). Gli Instant Avatar creano una copia virtuale sia dell’aspetto che della voce degli attori, replicando in tutto e per tutto le fattezze del video originale. Non è più necessario, quindi, dover selezionare l’avatar e la voce più adatti per il video in questione.

Creazione degli Instant Avatar Per la creazione degli Instant Avatar, come già anticipato, è stato usato un video a parte, registrato per ogni attore, realizzato seguendo le linee guida indicate dalla piattaforma per la creazione di un Instant Avatar:

- Durata di 2-5 minuti



Figura 10: Curva di equalizzazione dell’audio registrato.

- Frequenti pause tra una frase e l’altra
- Si può parlare di qualsiasi argomento, ma cercando di esagerare con le emozioni, così da donare espressività all’avatar
- Uso di movimenti delle braccia generici, evitando gesti di indicazione.

I video di riferimento registrati hanno una durata di circa 3 minuti. Caricato il video di riferimento, la piattaforma si occupa di tutto il resto, e in una decina di minuti l’avatar è pronto per essere utilizzato.

Creazione dei video

La creazione dei video fake è a sua volta semplificata. Creati i due Instant Avatar associati ai due attori, sono stati presi i due testi, scritti dal tesista della Cattolica Matteo Scarinzi, e sono stati caricati sulla piattaforma per la generazione dei video.

Come già visto nel Paragrafo 2.3.7, i testi sono stati separati in sezioni, così da introdurre naturalmente delle pause nel discorso generato. Inoltre, è stata aggiunta della punteggiatura aggiuntiva e sono state modificate alcune parole dei testi, esplicitandone l’accento, in modo da correggere alcuni problemi di intonazione e pronuncia che gli avatar hanno incontrato durante la generazione del testo. Ad esempio, la parola “prosodia” è stata scritta esplicitando l’accento, “prosodia”, così da correggerne la pronuncia, dal momento che la voce generata pronunciava la parola con l’accento sbagliato. La qualità e la precisione dell’intonazione del discorso generato dipendono fortemente dal discorso di riferimento che è stato utilizzato per la generazione dell’avatar. Seguendo l’esempio precedente, è evidente che i modelli di voce generati sono stati generati da un discorso che non conteneva la parola “prosodia” al suo interno,

lasciando al modello il compito di ricavarne la pronuncia, non garantendo ricavi quella corretta, rendendo necessario la specifica a mano in caso di errore.

Download del risultato

Sono stati realizzati un video fake per ogni attore (maschio e femmina), e per ogni attore un video per ognuno dei due testi realizzati, per un totale di quattro video fake, associati ai quattro video reali registrati. I video fake sono stati scaricati in formato 1080p e 25fps.

2.5 Pro e contro

In questa sezione analizziamo i pro e i contro dei due approcci valutati per la generazione dei video:

- Video generati da video scaricati dal web (seguendo la procedura descritta nel Paragrafo 2.3)
- Video generati con attori (come sopra 2.4)

2.5.1 Video generati da video scaricati dal web

Questo approccio è consistito nel trovare dei video validi da utilizzare come video reali sul web, e utilizzare gli avatar forniti dalla piattaforma HeyGen per creare dei video fake con lo stesso contenuto.

Pro

- Non è necessario scrivere dei testi appositi
- Non è necessario registrare da sé dei video, per cui non serve cercare attori, recuperare l'attrezzatura necessaria, e investire il tempo richiesto per la registrazione e la produzione dei video registrati
- Ottenimento dei video reali veloce.

Contro

- Non si ha controllo sulla qualità, argomento, lunghezza e complessità del contenuto esposto

- La ricerca dei video può richiedere tempo e fornire risultati insoddisfacenti a seconda dei criteri di selezione richiesti e la disponibilità della piattaforma di ricerca video utilizzata
- I video reali e fake presentano soggetti diversi, che dipendono dagli avatar forniti dalla piattaforma scelta, il che può influenzare l'esperienza visiva introducendo una variabile esterna indesiderata
- Creazione dei video fake più lenta, vista la necessità di selezione dell'avatar e della voce più adatti.

2.5.2 Video generati con attori

Questo approccio è consistito nella scrittura dei testi da esporre, la ricerca di due attori (maschio e femmina), la registrazione dei video reali, la creazione di avatar personalizzati e l'utilizzo dei testi scritti per la generazione dei video fake.

Pro

- I video reali e fake sono visivamente identici, eliminando qualsiasi variabile esterna, e l'unica differenza tra i due video è la loro natura (reale o fittizia)
- Si ha il totale controllo sul contenuto esposto, in termini di qualità, complessità, lunghezza e argomento, essendo questo scritto appositamente
- Generazione dei video fake molto più veloce, in quanto non serve selezionare l'avatar e la voce più adatta, essendo l'aspetto e la voce clonati dalla piattaforma di generazione dei video.

Contro

- Acquisizione dei video reali molto più lenta, in quanto richiede la scrittura dei contenuti, la ricerca di attori per la registrazione dei video, e la registrazione e post-produzione dei video stessi
- La qualità dei video fake realizzati dipende dalla qualità dei video registrati, in termini di qualità video/audio e in termini di qualità dell'avatar personalizzato creato.

2.5.3 La scelta

Visti i pro e i contro di ogni approccio, è stato scelto per la ricerca il secondo approccio, e quindi sono stati utilizzati i video realizzati con attori. Nonostante il tempo non indifferente richiesto dalla scrittura, acquisizione e produzione dei video reali, è stato valutato di estrema importanza avere lo stesso soggetto tra video reale e video fake, così da eliminare qualunque tipo di bias dovuto a soggetti diversi, e poter essere sicuri che eventuali risultati diversi ottenuti da i due tipi di video (reali o fake) possano essere causati solo da tale differenza, e non da fattori esterni.

Capitolo 3

Raccolta dati

In questo capitolo approfondiamo le modalità di acquisizione definite per l’acquisizione dei dati, a partire dallo sviluppo di un’interfaccia ad-hoc in Python per presentare i video previsti dal protocollo e somministrare tutti i questionari definiti, e la gestione della cattura e il salvataggio opportuno di tutti i dati e i segnali richiesti quali le espressioni facciali, la registrazione dello schermo, i dati di eye-tracking e i dati fisiologici.

3.1 Sviluppo dell’interfaccia

Per facilitare la fase di sperimentazione e raccolta dei dati è stato predisposto lo sviluppo di un’interfaccia, che guida il partecipante attraverso l’esperimento. L’interfaccia tiene traccia di tutte le acquisizioni effettuate, e in base al numero del partecipante, mostra i video previsti dal protocollo per tale partecipante, somministra i questionari definiti, e gestisce la cattura, il salvataggio e la sincronizzazione di tutti i segnali e i dati previsti, interfacciandosi con l’eye-tracker e il braccialetto Empatica¹. L’interfaccia è stata sviluppata in Python, per mantenere il programma cross-platform, non essendo sicuri a monte del sistema operativo su cui si sarebbe fatto girare l’applicativo in fase di sperimentazione. L’interfaccia è stata sviluppata utilizzando la libreria di interfaccia grafica PyQt5. Il codice prodotto è reperibile al seguente link: <https://github.com/cosfederico/Interfaccia-Tesi/>.

3.1.1 Design di sviluppo

L’interfaccia è stata disegnata per essere controllata autonomamente dal partecipante. Una volta eseguita la calibrazione dell’eye-tracker (che deve essere svolta manualmente

¹Insieme all’interfaccia è fornito uno script per il download automatico dei dati catturati dal braccialetto Empatica. Questo viene approfondito nel Paragrafo 3.3.5.

da un operatore), il partecipante è in grado di svolgere autonomamente l'esperimento, inserendo i propri dati e rispondendo alle domande proposte. Tutte le domande sono poste su una scala da 1 a 5, o in modalità a risposta multipla per le domande di comprensione, per cui il partecipante ha bisogno solo del mouse per interagire con l'interfaccia. È stato molto importante rendere l'interfaccia compatibile a un utilizzo a una sola mano, poiché su uno dei due polsi del partecipante (il polso della mano non dominante) è indossato il braccialetto Empatica per la misura del battito cardiaco e il livello di sudorazione. È caldamente consigliato da Empatica evitare di muovere tale polso per ottenere i risultati più accurati, per questo motivo è stata fatta tale scelta di design.

3.1.2 Selezione dei video

L'interfaccia gestisce autonomamente la selezione dei video da mostrare al partecipante corrente. Utilizzando lo storico delle acquisizioni eseguite sino a quel momento, l'interfaccia tiene conto del numero del partecipante attuale, e programmaticamente stabilisce i due video da somministrare a partire da tale numero, costruendo dinamicamente la matrice sperimentale presentata nella Sezione 1.1.3, di cui è visibile un esempio in Tabella 1.

Come già presentato in tale sezione, possiamo identificare univocamente un video tramite la valorizzazione di tre variabili binarie, quali *Contenuto*, *Tipo* e *Genere*, le quali rappresentano rispettivamente:

- *Contenuto*: il contenuto trattato, se A o B
- *Tipo*: il tipo di video, se Reale o Sintetico
- *Genere*: il genere del docente, se Maschile o Femminile

Andiamo a codificare le tre variabili indipendenti come variabili binarie con dominio $D_{Contenuto} = D_{Tipo} = D_{Genere} = \{0, 1\}$, dove:

$$\begin{aligned} Contenuto &= \begin{cases} 0 & \text{se il video contiene il Contenuto A} \\ 1 & \text{se il video contiene il Contenuto B} \end{cases} \\ Tipo &= \begin{cases} 0 & \text{se il video è Reale} \\ 1 & \text{se il video è Sintetico} \end{cases} \\ Genere &= \begin{cases} 0 & \text{se il genere del docente è Maschile} \\ 1 & \text{se il genere del docente è Femminile} \end{cases} \end{aligned}$$

Una valorizzazione di queste tre variabili rappresenta un possibile video da somministrare al nostro partecipante. La matrice è costruita tale per cui il valore delle variabili associate al secondo video è calcolato mediante la negazione logica dei valori delle variabili associate al primo video, questo per garantire il controbilanciamento. Per questo motivo, il calcolo è eseguito per determinare il primo video da essere visionato. I risultati ottenuti per il primo video sono poi negati per ottenere le proprietà del secondo video da mostrare.

Il calcolo è basato su tre operazioni binarie: trattando il numero del partecipante come un numero binario, sono estratti, in ordine, i tre bit meno significativi per stabilire i valori delle tre variabili binarie. Ad esempio, per il partecipante numero 6, questo codificato in binario risulta:

$$\begin{aligned} 6_{10} &= \underbrace{110} \\ Contenuto &= 0 \\ Tipe &= 1 \\ Genere &= 1 \end{aligned} \tag{1}$$

Nel caso del partecipante numero 14, sono ignorati i bit in eccesso:

$$\begin{aligned} 14_{10} &= 1 \underbrace{110} \\ Contenuto &= 0 \\ Tipe &= 1 \\ Genere &= 1 \end{aligned} \tag{2}$$

Come si può evincere, questo permette di stabilire immediatamente quali sono i video da somministrare, senza il bisogno di memorizzare esplicitamente la matrice sperimentale. Inoltre, l'esempio mostrato ci permette di fare un'altra osservazione interessante: i bit estratti per i partecipanti 6 e 14 sono gli stessi, per cui visioneranno gli stessi video, nello stesso ordine, rimarcando la natura periodica della matrice sperimentale.

Il calcolo esplicitato, per un generico partecipante di numero $ID_{partecipante}$, può essere scritto come:

$$\begin{aligned}
 \text{Contenuto} &= ID_{\text{partecipante}} \bmod 2 \\
 ID_{\text{partecipante}} &\leftarrow \lfloor ID_{\text{partecipante}} / n \rfloor \\
 \text{Tipo} &= ID_{\text{partecipante}} \bmod 2 \\
 ID_{\text{partecipante}} &\leftarrow \lfloor ID_{\text{partecipante}} / n \rfloor \\
 \text{Genere} &= ID_{\text{partecipante}} \bmod 2
 \end{aligned} \tag{3}$$

Si può osservare che non vi è in realtà alcun limite al numero di contenuti diversi che si potrebbero trattare, per cui la variabile contenuto non è limitata a essere binaria. In tal caso il calcolo per il contenuto da mostrare diventerebbe, più genericamente:

$$\begin{aligned}
 \text{Contenuto} &= ID_{\text{partecipante}} \bmod k \\
 ID_{\text{partecipante}} &\leftarrow \lfloor ID_{\text{partecipante}} / k \rfloor \\
 \dots
 \end{aligned} \tag{4}$$

dove k è il numero di contenuti diversi trattati nello studio.

3.1.3 Fase di preparazione

L’interfaccia è stata arricchita di alcune funzionalità a supporto dello sperimentatore. Prima di cominciare con l’esperimento, viene aperta una finestra di preparazione, dove è possibile selezionare tra tutte le webcam attive sul sistema quale utilizzare per la registrazione del video, ed è presente un pulsante per effettuare una prova del sistema audio, per assicurarsi di avere selezionato l’uscita audio corretta (Figura 11a). A seguire, è visualizzata una finestra di anteprima della webcam selezionata. In sovrapposizione all’anteprima della webcam è posta una figura stilizzata per indicare il posizionamento ottimale della persona e del volto nell’inquadratura, per l’analisi ottimale delle espressioni facciali (Figura 11b). Terminata la fase di preparazione, viene lanciato il programma principale, dove il partecipante è accolto nella schermata di introduzione (Figura 12), dove è informato sui dettagli dell’esperimento, quali durata dei video e dell’esperimento, dati e segnali raccolti, e accorgimenti sull’utilizzo del mouse. Qui, il partecipante ha la possibilità di esprimere il suo consenso informato, tramite checkbox, e avviare l’esperimento, altrimenti uscire e abbandonare l’esperimento. L’avvio della registrazione di tutti i dati e segnali avviene solo una volta che il partecipante ha espresso il suo consenso e premuto il pulsante “Inizia”. A partire da quel momento, l’esperimento è ufficialmente iniziato.

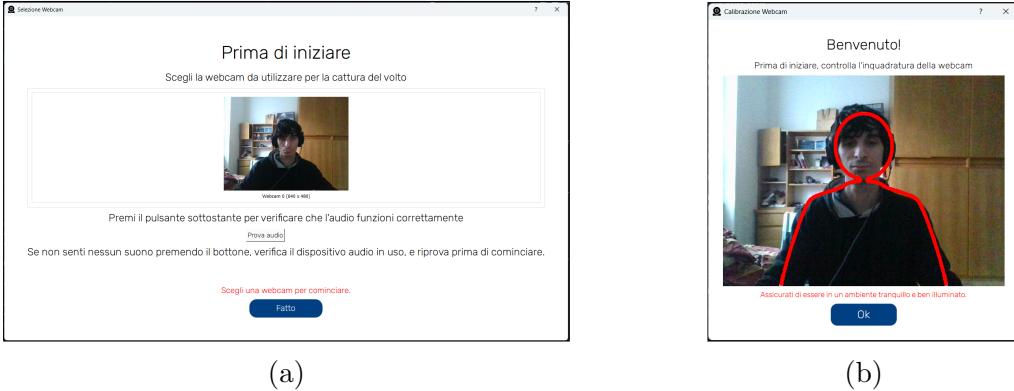


Figura 11: Schermate di preparazione dell’interfaccia sviluppata: (a) Schermata di selezione della webcam. (b) Schermata di anteprima della webcam, con figura in sovra-impressione per il posizionamento ottimale del volto.

3.1.4 Raccolta dei timestamp

L’interfaccia tiene traccia di tutti i timestamp degli eventi principali, per permettere la sincronizzazione dei dati e dei segnali raccolti, e per poter associare i dati agli eventi di interazione del partecipante con l’interfaccia. L’interfaccia è stata disegnata a pagine: ogni volta che il partecipante passa alla pagina successiva viene salvato il timestamp di tale evento. In questo modo, sono salvati i timestamp di inizio sessione, fine sessione, inizio e fine di ogni video visualizzato, e i timestamp di risposta ad ogni domanda dei questionari presentati. I questionari sono stati disegnati in modo da presentare una domanda alla volta per pagina²: quando il partecipante passa alla domanda successiva, viene salvato il timestamp di tale evento. Questo ci permette non solo di sapere quando il partecipante è passato alla domanda successiva, ma anche quanto tempo ha impiegato per rispondere alla domanda proposta, sottraendo il timestamp salvato con il precedente. Trattandosi di domande a risposta chiusa (da 1 a 5, o a risposta multipla) non ci si deve preoccupare del tempo di digitazione della risposta, in quanto per rispondere basta un click. I timestamp sono salvati in UNIX-time, rispetto al fuso orario UTC (Coordinated Universal Time).

3.1.5 Registrazione della webcam/dello schermo

Durante tutta la durata dell’esperimento, l’interfaccia si occupa della registrazione della webcam per l’estrazione delle espressioni facciali, e della registrazione dello

²Fatta eccezione per il questionario di autovalutazione dello stato emotivo (Positive and Negative Affect Schedule, PANAS), il quale è presentato in una pagina sola, e di questo è salvato solo il timestamp di inizio e fine di tutto il questionario.

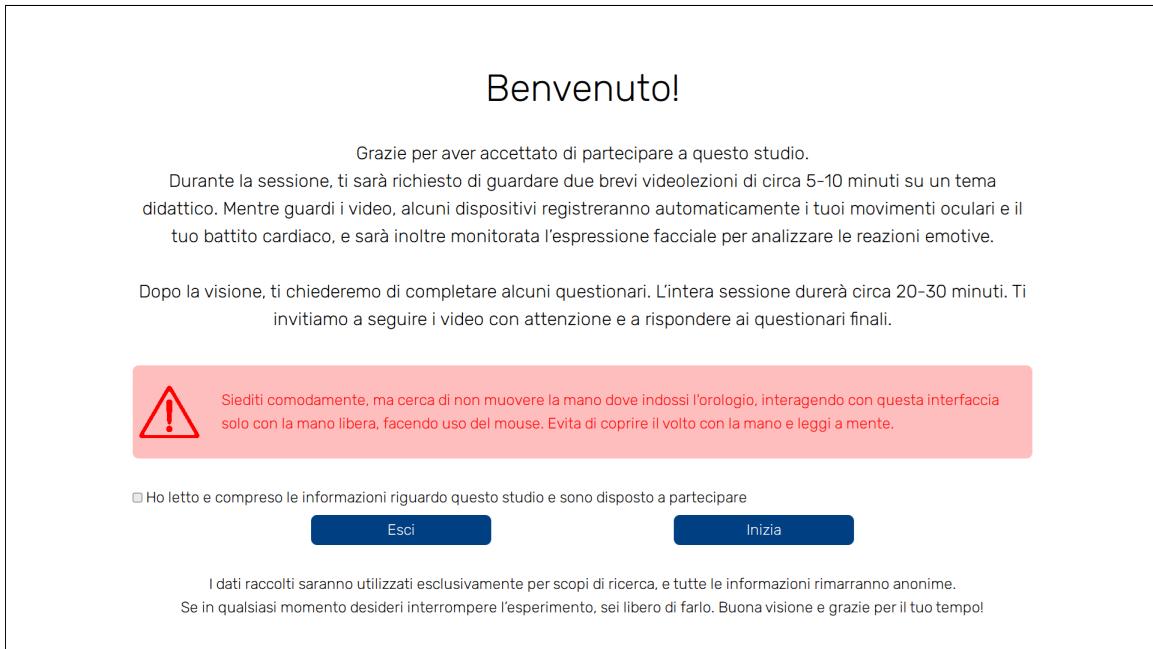


Figura 12: Schermata di benvenuto dell'interfaccia sviluppata.

schermo.

Registrazione della webcam La registrazione del volto è effettuata tramite una webcam esterna, posta sotto al monitor sul quale sono presentati gli stimoli, e come già detto in fase di avvio l'interfaccia permette di selezionare quale webcam utilizzare per la registrazione del volto. Una volta selezionata, è mostrata allo sperimentatore una schermata con le istruzioni necessarie per il posizionamento ottimale della webcam (Figura 11b).

Registrazione della webcam La registrazione dello schermo è necessaria per l'utilizzo dei dati di eye-tracking, in quanto rappresenta gli stimoli presentati. Combinando i dati di eye-tracking e la registrazione dello schermo è possibile analizzare il movimento dello sguardo e le fissazioni, associando ad ogni sguardo e fissazione lo stimolo sullo schermo che lo ha stimolato.

Le registrazioni sono eseguite su due processi paralleli al processo principale di generazione della GUI. Questo permette di non rallentare l'interfaccia grafica e permette una registrazione di entrambi i video più fluida, riducendo di molto il rischio di fotogrammi persi.

3.1.6 Controllo dell'eye-tracker

L'interfaccia si occupa anche del controllo dell'eye-tracker, e quindi dell'acquisizione e del salvataggio dei dati di eye-tracking. Questo è collegato al computer che esegue l'interfaccia tramite cavo ethernet, ed è controllato utilizzando le librerie di API fornite dal produttore dell'eye-tracker.

A inizio esperimento, l'interfaccia effettua il collegamento con l'eye-tracker e avvia la procedura di calibrazione. La calibrazione è svolta manualmente dallo sperimentatore, il quale guida il partecipante attraverso il processo, e, dovesse essere necessario, riesegue la calibrazione fino ad ottenere dei risultati abbastanza buoni per una buon tracciamento dello sguardo. Quando la calibrazione è conclusa, il controllo torna all'interfaccia (Figura 12), da dove il partecipante può far iniziare l'esperimento.

Quando il partecipante preme il pulsante per iniziare l'esperimento, l'interfaccia avvia la cattura dei dati di eye-tracking, gestendo il salvataggio del timestamp di inizio cattura, per permettere poi la sincronizzazione dei dati di eye-tracking con tutti gli altri dati.

3.2 Salvataggio dei dati raccolti

In questa sezione andiamo ad analizzare brevemente come tutti i dati raccolti dall'interfaccia vengono salvati, per essere poi pronti per la fase di analisi.

3.2.1 Dati demografici e questionari

I dati misti catturati dall'interfaccia, quali i dati demografici del partecipante (età, genere, nazionalità, livello di istruzione), le risposte ai questionari, e tutti i timestamp registrati, sono salvati in un formato .csv. Il file .csv presenta una sola riga, con campi:

- ID del partecipante
- Dati demografici (età, genere, nazionalità, livello di istruzione)
- Timestamp di inizio, fine sessione
- Timestamp di inizio, fine video, natura e genere dei video mostrati
- Per ogni domanda la risposta selezionata e la domanda associata + timestamp
- Risposte ai questionari PANAS, con punteggio ottenuto + timestamp
- Risposte alle domande a risposta multipla, salvando la domanda chiesta, la risposta segnata, e se questa era corretta o meno + timestamp.

3.2.2 Video

Per l'estrazione delle espressioni facciali (registrazione della webcam) e l'analisi degli stimoli visivi (registrazione dello schermo) non è necessario l'utilizzo del segnale RAW della webcam, per cui le registrazioni sono salvate in formato .mp4 compresso, per ridurre lo spazio richiesto per memorizzare i dati salvati.

3.2.3 Eye-tracking

Come già detto, il sistema di eye-tracking è controllato internamente dall'interfaccia, utilizzando le API fornite da SR Research. A fine esperimento, l'interfaccia si occupa di interrompere l'acquisizione dei dati di eye-tracking e scaricare i dati registrati. I dati sono trasferiti via Ethernet, e sono scaricati in un formato proprietario (.edf, Eyelink Data Format), ma Research SR fornisce un programma da linea di comando per convertire tale formato in formato di testo (.txt). L'interfaccia si occupa quindi di scaricare i dati e convertirli in formato .txt. Il file .txt prodotto presenta una riga per campione, con posizione x, y dello sguardo sullo schermo, e diametro della pupilla, in numero di pixel. Per l'analisi dei dati è stato sviluppato uno script per la conversione dei dati dal formato .txt al formato .csv.

3.3 Dati fisiologici

Dedichiamo una breve sezione a parte per i dati fisiologici, i quali non sono controllati dall'interfaccia direttamente, bensì tramite applicazione via smartphone.

3.3.1 Setup

Per l'acquisizione dei segnali fisiologici è stato scelto l'utilizzo di un braccialetto Empatica EmbracePlus. Il braccialetto, indossabile al polso come un normale orologio, permette l'acquisizione non invasiva dei segnali che ci interessano, quali il battito cardiaco e il livello di sudorazione della pelle, così come anche molto altri segnali. Il braccialetto non presenta delle API per il controllo diretto, ma è pilotato via un'applicazione via smartphone. L'applicazione in questione è CareLab, prodotta da Empatica, e disponibile sia per iOS che per Android. Per potersi collegare all'orologio è necessario scaricare l'applicazione, effettuare il login, e solo allora il telefono è in grado di collegarsi al braccialetto tramite Bluetooth.

La generazione delle credenziali per accedere all'app è eseguita tramite un portale dedicato, fornito da Empatica a tutti i suoi clienti (<https://auth.carelab.empatica.com/>). Da questo portale, lo sperimentatore crea un "Partecipante": un

partecipante non è altro che una coppia di credenziali (username e password), utilizzabili per effettuare l'accesso all'app. Effettuato l'accesso all'app, lo smartphone si collega all'orologio, mostrando le istruzioni per indossarlo correttamente, e avvia immediatamente l'acquisizione dei dati. Per la gestione dei dati acquisiti, è stato scelto di utilizzare un solo partecipante per raccogliere i dati associati alla sperimentazione. Il portale è condiviso tra tutti i dispositivi di un associazione, ed essendo il braccialetto utilizzato di proprietà dell'Università, è bene concentrare tutti i dati in un posto solo. Per la distinzione delle catture è stato utilizzato il tempo di inizio dell'esperimento, univoco per ogni acquisizione.

3.3.2 Calibrazione

Il braccialetto non prevede una particolare fase di calibrazione. Una volta collegato via app, è richiesto massimo un minuto di configurazione eseguita automaticamente. Detto questo, come consigliato dalle guida di Empatica (<https://support.empatica.com/hc/en-us/articles/203621955-What-should-I-know-to-use-EDA-data-in-my-experiment>), per la cattura e l'utilizzo del livello di sudorazione della pelle (o ElectroDermal Activity, EDA), è consigliato indossare l'orologio in condizioni di riposo per circa 10-15 minuti prima dell'inizio dell'esperimento, per permettere alla pelle del partecipante di adattarsi al contatto con il sensore, e permettere l'acquisizione di un buon segnale di sudorazione.

3.3.3 Acquisizione dei segnali

Dopo la fase di calibrazione, il braccialetto inizia subito a catturare dati. L'acquisizione è quindi automatica e asincrona. I dati vengono trasferiti continuamente sul telefono e sincronizzati periodicamente su un server Amazon AWS. Per questo motivo, è importante che la cattura dei dati fisiologici sia avviata e terminata rispettivamente prima e dopo che l'esperimento sia iniziato e finito, così da permettere di sincronizzare i segnali fisiologici con i dati raccolti dell'interfaccia.

3.3.4 Salvataggio dei dati

Una volta terminato l'esperimento, è presente sull'app un pulsante per interrompere l'acquisizione, il quale arresta l'acquisizione dei dati, trasferisce i dati non ancora sincronizzati, li carica sul cloud online e spegne l'orologio. A seguito di questa operazione, i dati sono al sicuro sul cloud, e sono scaricabili in qualsiasi momento, accedendo al cloud tramite le credenziali AWS, fornite da Empatica sul portale indicato. Questa

soluzione è stata valutata come molto utile, in quanto, posto che l'orologio sia stato configurato correttamente, il salvataggio dei dati è automatico, non dovendosene preoccupare durante la sperimentazione.

3.3.5 Download dei dati

Per il download dei dati dal cloud è stato sviluppato uno script in Python, che utilizzando l'ID del partecipante, l'ora dell'acquisizione e i dati di login al server AWS, scarica automaticamente i dati associati alla acquisizione e li sincronizza con la cattura, ritagliando i dati che ci servono. Questa operazione verrà approfondita nel Capitolo 4 sull'elaborazione dei segnali. Lo script si occupa anche di selezionare i segnali che si vogliono scaricare. Il braccialetto Empatica in realtà è in grado di catturare un gran numero di segnali, e in condizioni normali li cattura tutti. Questi segnali sono:

- Accelerometro (x, y, z)
- Giroscopio (x, y, z)
- Temperatura della pelle ($^{\circ}\text{C}$)
- **ElectroDermal Activity** (μS)
- Passi
- **Blood Volume Pulse (BVP)** ($n\text{W}$)
- **Systolic Peaks**

Tra questi sono indicati in grassetto i dati che vengono scaricati, quali: i dati di picchi sistolici (systolic peaks), per il battito cardiaco, il livello di sudorazione della pelle (ElectroDermal Activity, EDA), e anche il segnale di BVP (volume di pressione nel sangue), il quale è sempre associato al battito cardiaco, e potrebbe tornare utile in fase di analisi. In verità, il braccialetto produce anche una serie di segnali aggregati, ovvero segnali più complessi come il battito cardiaco o il ritmo respiratorio, ma questi segnali sono aggregati per minuto, ovvero vi è un campione ogni minuto. Questo rende l'utilizzo di questi dati aggregati non applicabile per la nostra ricerca, la quale richiede una precisione più fine³. I segnali indicati sono i così detti segnali RAW. Nel Capitolo 4 verrà approfondita l'analisi di questi segnali, e come questi segnali vengono utilizzati, ad esempio come il segnale di picchi sistolici viene utilizzato per ricostruire il segnale di battito cardiaco in forma non aggregata.

³Volendo analizzare il comportamento di questi segnali durante la visione di video, dalla durata media di 6-7 minuti, un dato aggregato per minuto sarebbe molto poco utile, il quale presenterebbe nel caso migliore 7 campioni per video.

Capitolo 4

Analisi dei dati acquisiti

In questo capitolo affrontiamo l'analisi di tutti i dati acquisiti, a partire da come questi sono stati preparati all'analisi, ovvero la suddivisione in due popolazioni: i dati catturati durante la visione di un video reale e i dati catturati durante la visione di un video sintetico. In seguito, affrontiamo l'analisi statistica effettuata per ogni tipo di dato raccolto (eye-tracking, espressioni facciali, dati fisiologici, risposte ai questionari, tempi di risposta), con lo scopo di valutare l'esperienza di visione dei video presentati, indagando la possibile presenza di differenze tra i gruppi. Infine, i risultati ottenuti vengono valutati, per trarre le dovute conclusioni.

4.1 Preparazione dei dati

4.1.1 Suddivisione in gruppi

Incominciamo questo percorso finale dalla preparazione dei dati alla fase di analisi. Sono state effettuate un totale di 20 acquisizioni. Per la preparazione all'analisi, i dati sono stati suddivisi in due gruppi: il gruppo di dati raccolti durante e a seguito (per i questionari) della visione di un video reale, che chiamiamo “real”, e il gruppo di dati raccolti durante e a seguito della visione di un video di natura sintetica, che chiamiamo “fake”. Questo produce 2 popolazioni con 20 campioni appaiati. L'analisi è stata strutturata investigando la presenza di eventuali differenze tra le due popolazioni tramite analisi qualitativa e test di ipotesi. Sono state messe a confronto le distribuzioni dei dati raccolti, così come eventuali feature particolari, proprie del dominio dei dati specifici.

I dati raccolti rientrano nelle seguenti categorie:

- Dati fisiologici: Battito cardiaco (HR), ElectroDermal Activity (EDA)
- Espressioni facciali: Espressioni discrete, AUs

- Eye-tracking: Fissazioni, saccadi, battiti di ciglia
- Questionari: Risposte, tempi di risposta

4.2 Espressioni facciali

Iniziamo la nostra analisi dai dati di espressione facciale acquisiti durante la fase di sperimentazione. I dati sono stati suddivisi in due gruppi: i dati acquisiti durante la visione di un video reale (“real”), e i dati acquisiti durante la visione di un video sintetico (“fake”). I dati sono stati estratti utilizzando il package `py-feat` [8], a partire dalle registrazioni del volto dei partecipanti effettuate tramite webcam durante la visione dei video¹. Questo package fornisce: dati sull’identità, le espressioni discrete (ad es. felicità, tristezza, rabbia), i movimenti muscolari facciali (Action Units, AUs) e i landmark facciali. Inoltre, fornisce una serie di metodi per il pre-processing, l’analisi e la visualizzazione di dati di espressioni facciali. Per questo studio, sono stati analizzati i dati relativi alle espressioni discrete e le AUs. Innanzitutto, sono state confrontate l’espressione media, minima, e più estrema tra i due gruppi. In seguito, sono stati confrontati i valori massimi e medi delle espressioni discrete assunte durante la visione dei video. Utilizzando gli strumenti forniti dal package `py-feat`, sono stati condotti dei *t*-test indipendenti per confrontare le distribuzioni delle espressioni discrete. È stato costruito e valutato un classificatore, a partire da una serie di combinazioni (espressioni, espressioni + posa, AUs + posa, AUs, posa), per identificare quale tipo di video si stesse guardando, a partire dai soli dati di espressione facciale. Infine, sono state studiate le Action Units (AUs), attraverso un confronto grafico, utilizzando la tecnica di proiezione su due dimensioni UMAP², per investigare la possibile presenza di cluster distinti tra i due gruppi.

4.2.1 Espressione media, minore, più estrema

Sono state estratte, tra i due gruppi, l’espressione facciale media, minima e massima assunta durante la visione dei due tipi di video, in altre parole i valori medi, massimi e minimi delle AUs. Le espressioni sono state confrontate tramite test di ipotesi, utilizzando il package `autorank`, confrontando i valori medi, massimi e minimi di ogni AU tra i due gruppi. L’analisi è stata condotta per 2 popolazioni, “real” e “fake”, con 20 campioni appaiati (i 20 valori di AUs). È stato utilizzato come livello di significatività $\alpha=0.050$. A seguito di un test di normalità di Shapiro-Wilk, i valori medi e massimi

¹Sono stati estratti e processati 1 frame al secondo.

²Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) è una tecnica di riduzione di dimensioni di un vettore, che permette di ridurre un vettore di qualsiasi dimensione a un vettore a due dimensioni, permettendone una visualizzazione grafica più naturale.

Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
Espress. media	0.278	0.168	0.280	0.168	0.564	inconclusive
Espress. massima	0.882	0.112	0.895	0.112	0.321	inconclusive
Espress. minima	0.025	0.025	0.015	0.015	0.074	inconclusive

Tabella 2: Risultati dell’analisi dell’espressione facciale media, massima e minima tramite test di ipotesi, condotti utilizzando il package `autorank`. Sono riportati i valori di media (M) e deviazione standard (SD) per i valori valutati come normali (righe evidenziate in grigio). Sono riportati i valori di mediana (MD) e deviazione mediana assoluta (MAD) per i valori valutati come non normali (righe rimanenti).

sono stati valutati come normali (*p*-value minimo osservato=0.056), pertanto è stato effettuato un *t*-test per il confronto, i valori minimi sono stati valutati come non normali ($p_{real} = p_{fake} = 0.000$), pertanto questi sono stati confrontati tramite test dei ranghi con segno di Wilcoxon. Il confronto non ha mostrato differenze significative né per l’espressione media ($p = 0.564$), né per l’espressione massima ($p = 0.321$), né per l’espressione minima ($p = 0.074$), sebbene i risultati relativi all’espressione minima siano molto vicini alla significatività. In Figura 13 sono visualizzate le espressioni studiate. Anche confrontando le immagini, vi sono minime differenze per le espressioni facciali medie e massime, e vi è una intensità maggiore dell’AU17 nell’espressione facciale minima associata ai video reali. La AU è denominata “Chin Raiser”, ed è associata all’espressione discreta del disgusto³. Questa differenza è, molto probabilmente, responsabile della forte differenza tra i *p*-value trovati. Nonostante ciò, in base ai dati a disposizione e ai test condotti, questa differenza non è abbastanza grande da risultare statisticamente significativa. In Tabella 2 sono riassunti i risultati ottenuti.

4.2.2 Espressioni discrete

Passiamo ora all’analisi delle espressioni discrete, indagando se vi sono espressioni più predominanti all’interno di un gruppo piuttosto che di un altro. Per espressioni discrete intendiamo una serie di espressioni facciali predeterminate: felicità, tristezza, sorpresa, paura, disgusto, rabbia, neutrale. Per ogni espressione estratta, il package `py-feat` fornisce una percentuale di affinità dell’espressione corrente con ognuna di queste espressioni discrete. L’espressione discreta predominante in tutto lo studio è l’espressione neutrale ($M=77.13\%$ per i video reali, $M=75.64\%$ per i video sintetici), seguita dall’espressione di tristezza ($M=16.7\%$, $M=17.2\%$) e di rabbia ($M=3.26\%$, $M=3.88\%$). Analizzando i valori medi, è stata osservata una maggiore espressività

³Fonte: https://py-feat.org/pages/au_reference.html

Espressioni	(%)	Reale		Sintetico		<i>p</i> -value	Decisione
		MD	MAD	MD	MAD		
Medie							
Rabbia	0.92	0.64	0.83	0.54	0.156	inconclusive	
Disgusto	0.11	0.09	0.16	0.13	0.007	real < fake	
Paura	0.18	0.13	0.17	0.17	0.885	inconclusive	
Felicità	0.24	0.13	0.23	0.15	0.551	inconclusive	
Tristezza	4.55	4.19	4.85	4.00	0.435	inconclusive	
Sorpresa	0.41	0.40	0.68	0.64	0.364	inconclusive	
Neutrale	84.54	11.78	86.46	10.29	0.565	inconclusive	
Massime							
Rabbia	6.00	4.32	6.47	4.60	0.676	inconclusive	
Disgusto	2.50	1.94	3.86	3.55	0.035	real < fake	
Paura	1.70	1.21	4.22	3.35	0.464	inconclusive	
Felicità	7.70	6.15	12.07	9.568	0.420	inconclusive	
Tristezza	39.13	32.82	35.22	29.03	0.899	inconclusive	
Sorpresa	11.58	11.50	11.19	11.11	0.364	inconclusive	
Neutrale	97.21	1.76	97.53	1.85	0.324	inconclusive	

Tabella 3: Confronto dei valori medi e massimi delle espressioni discrete per partecipante tra i gruppi, tramite test dei ranghi con segno di Wilcoxon. Sono riportati i valori di mediana (MD) e deviazione mediana assoluta (MAD). L’analisi è stata svolta utilizzando il package **autorank**.

durante la visione di video sintetici, riportando un’intensità minore dell’espressione neutrale ($75.64\% < 77.13\%$), e un intensità maggiore per tutte le altre espressioni discrete. Per effettuare un confronto appaiato, aggreghiamo questi dati per partecipante, e calcoliamo una media dei valori di ogni espressione discreta per ogni partecipante tra i gruppi. Otteniamo così, per ogni espressione discreta, 2 popolazioni con 20 campioni appaiati. È stato effettuato un test di normalità di Shapiro-Wilk per ogni popolazione, il quale ha valutato tutte le popolazioni come non normali ($p \leq 0.003$), pertanto, è stato utilizzato il test dei ranghi di Wilcoxon per il confronto. Sono state trovate differenze significative tra le distribuzioni dei valori medi dell’espressione discreta del “Disgusto” (Reale=0.15%, AI=0.30%, $p = 0.007$), riportando un valore medio mediamente il doppio più elevato durante la visione di un video con presentatore sintetico. In base ai dati a disposizione, le altre espressioni discrete non hanno mostrato differenze statisticamente significative ($p \geq 0.156$), in termini di tendenza centrale dei valori medi. Confrontando i valori massimi, è stato rilevato un picco massimo del “Disgusto”, seguito dall’espressione della “Paura” (Reale=74.47%,

Espressione (%)	M _{Reale}	M _{AI}	<i>t</i> _{stat}	<i>p</i> -value
Rabbia	0.033	0.039	-4.18	$2.92 \cdot 10^{-5}$
Disgusto	0.002	0.003	-5.49	$4.1 \cdot 10^{-8}$
Paura	0.005	0.006	-3.25	0.00117
Felicità	0.005	0.005	-1.46	0.143
Tristezza	0.168	0.172	-1.25	0.212
Sorpresa	0.018	0.018	-0.85	0.397
Neutrale	0.771	0.756	3.72	0.000201

Tabella 4: Risultati dell’analisi tramite *t*-test indipendente tra i valori di espressioni discrete tra i gruppi, svolta usando il package `py-feat`. Sono evidenziate le righe per cui sono state trovate differenze statisticamente significative.

AI=96.41%). Ripetendo la stessa analisi effettuata per le espressioni medie, i risultati sono analoghi, riportando differenze statisticamente significative tra le distribuzioni dei valori massimi dell’espressione del “Disgusto” ($p = 0.035$), raggiungendo picchi massimi maggiori durante la visione di video sintetici. I risultati delle analisi condotte sono riportati in Tabella 3.

Per concludere l’analisi delle espressioni discrete, sono stati utilizzati gli strumenti messi a disposizione dal package `py-feat` per condurre dei *t*-test tra campioni indipendenti per confrontare le distribuzioni globali dei valori delle espressioni discrete tra i due gruppi. I test pongono come ipotesi nulla che le due popolazioni confrontate appartengano alla stessa distribuzione, in particolar modo con la stessa media. L’analisi è stata condotta, per ogni espressione discreta, tra due popolazioni, “real” ($N=8437$) e “fake” (8304). È stato utilizzato come livello di significatività $\alpha=0.050$. I test condotti hanno riportato differenze statisticamente significative tra i valori medi delle espressioni discrete della “Rabbia” ($t = -4.18, p = 2.92 \cdot 10^{-5}$), del “Disgusto” ($t = -5.49, p = 4.1 \cdot 10^{-8}$), della “Paura” ($t = -3.25, p = 0.00117$) e per l’espressione neutrale ($t = 3.72, p = 0.000201$), riportando per le prime tre, un valore medio significativamente maggiore per la popolazione “fake” ($t < 0$), riportando invece per l’espressione neutrale un valore medio significativamente maggiore per la popolazione “real” ($p > 0$). In Tabella 4 sono riportati i risultati trovati. In Figura 14 sono riportati i boxplot delle distribuzioni studiate. Analizzando qualitativamente i grafici, effettivamente risaltano subito all’occhio le espressioni discrete della “Rabbia”, del “Disgusto” e della “Paura”, caratterizzate da delle distribuzioni più dilatate, e con picchi più alti, durante la visione di video sintetici. Al contrario, si può osservare una distribuzione più spostata verso destra durante la visione dei video reali per l’espressione della “Felicità” e dell’espressione “Neutrale”. Risultano, infine, molto simili le distribuzioni delle espressioni della “Tristezza”, della “Sorpresa”.

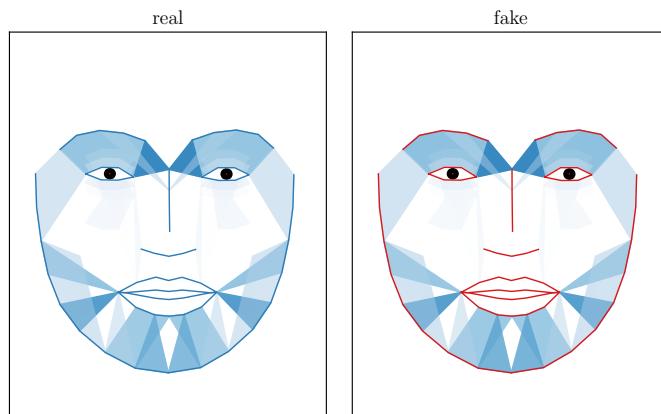
I risultati trovati indicano una chiara differenza, riportando una maggiore predominanza delle espressioni di “Rabbia”, “Paura” e in particolar modo del “Disgusto”, in termini di valore medio e picchi massimi, durante la visione di video con un presentatore sintetico, nonostante ricordiamo che parliamo sempre di piccole percentuali. È bene sottolineare come le analisi appena condotte trattano puramente di espressioni facciali, non intendiamo assumere che vi sia alcuna correlazione tra le espressioni assunte e lo stato emotivo dei partecipanti durante la visione. In altre parole, questi risultati non indicano che i partecipanti abbiano provato a livello emotivo emozioni diverse a seguito della visione dei due tipi di video, parliamo strettamente di espressioni facciali.

4.2.3 Creazione di un classificatore

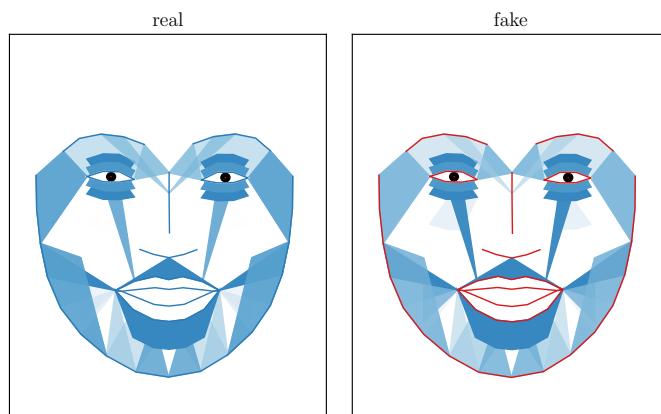
Utilizzando gli strumenti forniti dal package `py-feat`, sono stati addestrati e valutati cinque classificatori, con lo scopo di distinguere tra le popolazioni “real” e “fake”, utilizzando sottoinsiemi diversi dei dati raccolti. I classificatori sono stati addestrati a partire dai sottoinsiemi: espressioni discrete, espressioni discrete + posa, AUs + posa, AUs, posa. In Tabella 15a sono riportati i valori di precisione dei classificatori ottenuti, e in Figura 15b questi risultati sono visualizzati tramite grafico a barre. I risultati sono molto interessanti: tutti i classificatori hanno ottenuto una precisione media del 52.04% ($SD=4.60\%$). In altre parole, basandosi soltanto sui dati di espressione facciale acquisiti durante il nostro esperimento, cercare di identificare che tipo di video il soggetto stesse guardando a partire dai dati di espressione facciale sarebbe tanto preciso quanto lanciare una moneta e tirare a indovinare. Questo è un risultato molto forte, in quanto evidenzia come, complessivamente, in base ai dati a disposizioni, non vi sono differenze significative nel comportamento delle espressioni facciali tra un soggetto che visiona una lezione con un presentatore reale e un soggetto che visiona una lezione con un presentatore sintetico.

4.2.4 Confronto delle AUs con UMAP

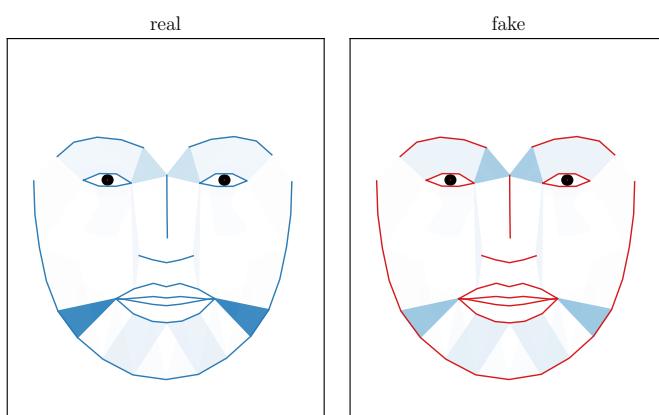
L’ultima analisi fatta sui dati di espressione facciale è stata confrontare graficamente le AUs tra le due popolazioni tramite una proiezione su due dimensioni, utilizzando una tecnica chiamata Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). Per UMAP si intende una tecnica di riduzione delle dimensioni di un vettore, utilizzata per la visualizzazione su due dimensioni di vettori multidimensionali [17], come le nostre AUs, le quali consistuiscono un vettore a 20 dimensioni (20 AUs). L’operazione di riduzione è stata realizzata con il package UMAP. La proiezione è mostrata in Figura 16. I risultati vanno interpretati qualitativamente: se vi fossero differenze significative tra le due popolazioni, sarebbero identificabili dei



(a) Espressione facciale media



(b) Espressione facciale massima



(c) Espressione facciale minima

Figura 13: Confronto delle espressioni facciali assunte durante la visione di un video reale piuttosto che sintetico: (a) espressione media, (b) espressione massima, (c) espressione minima.

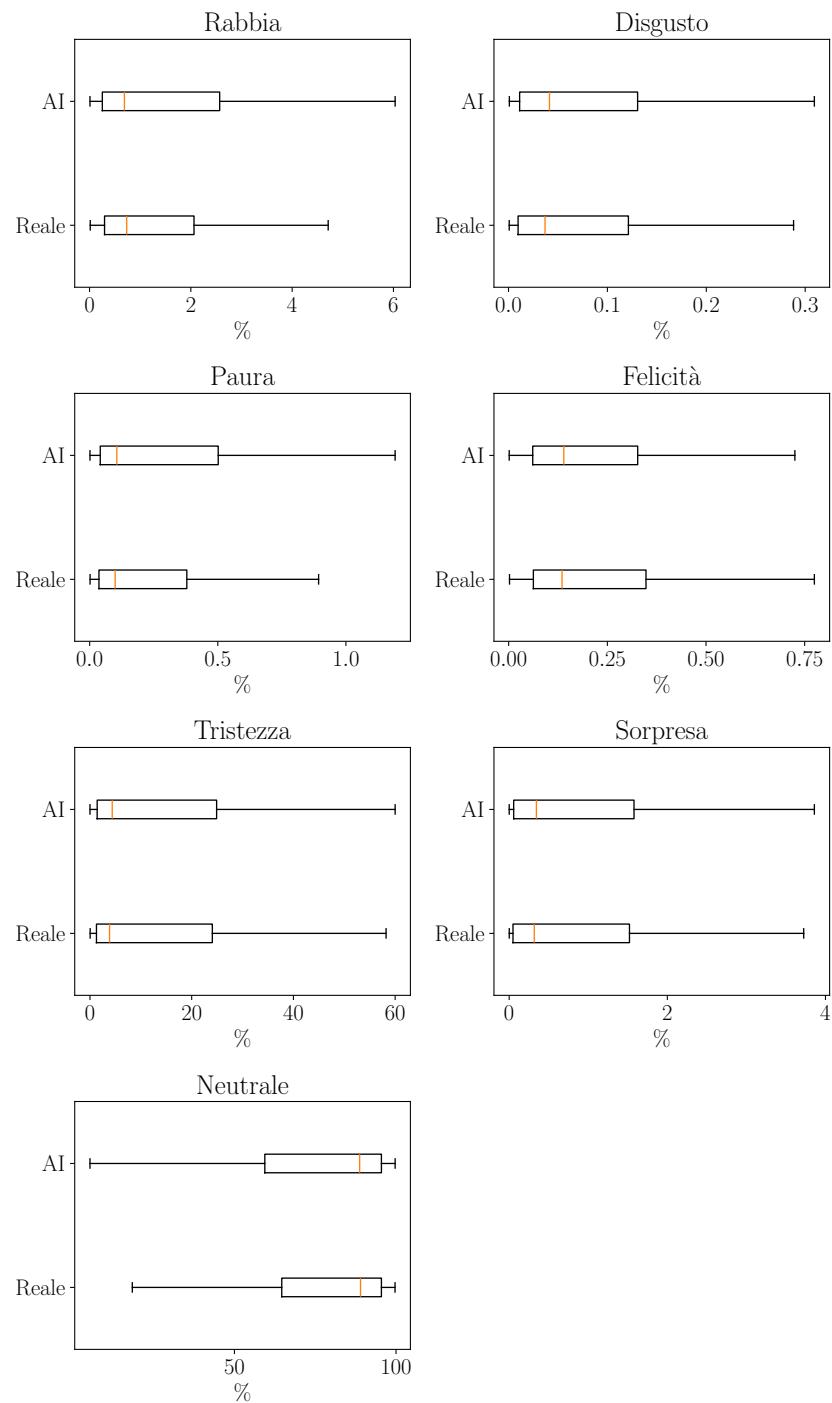


Figura 14: Boxplots delle distribuzioni dei valori di espressioni discrete tra i gruppi.

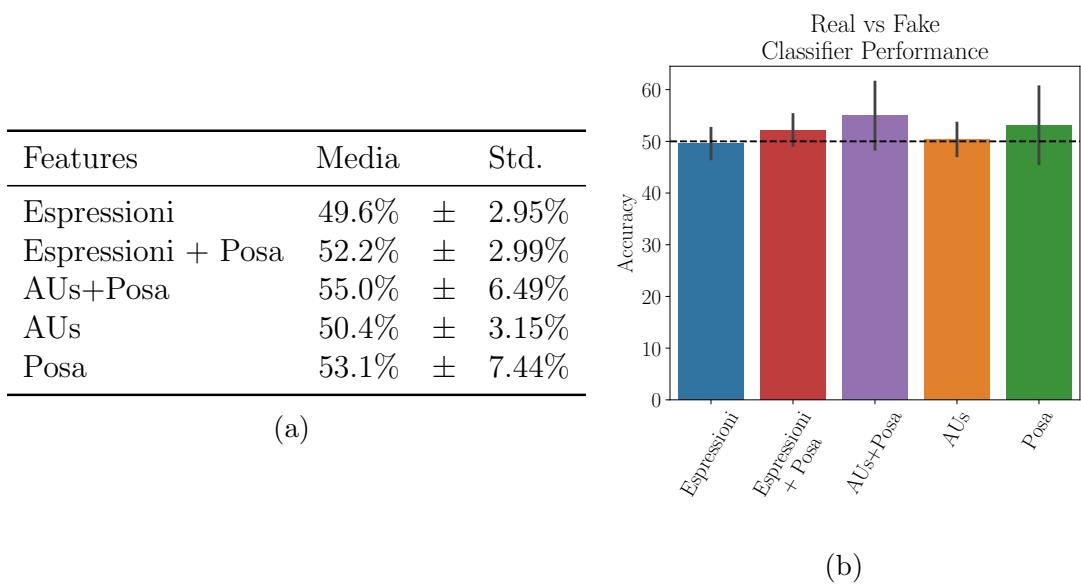


Figura 15: Risultati del training di classificatori per distinguere tra video reale e sintetico a partire da: Espressioni, espressioni + posa, AUs + posa, AUs e posa. (a) Risultati in forma tabellare, (b) visualizzazione tramite grafico a barre.

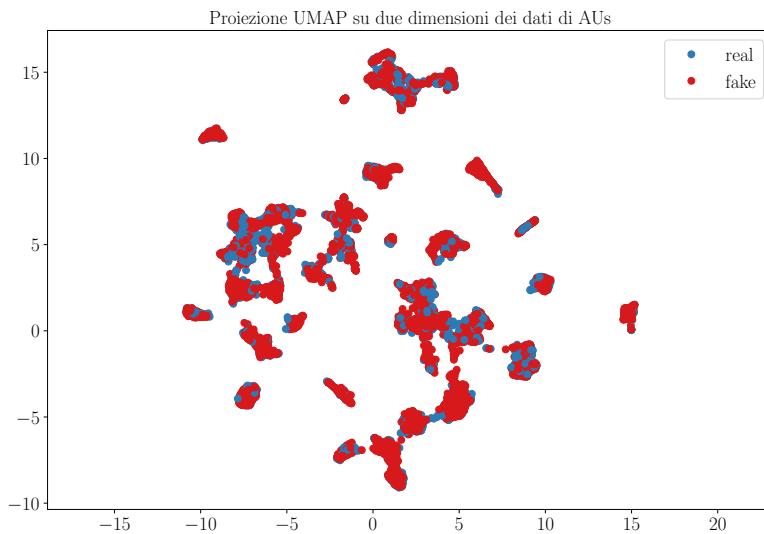


Figura 16: Proiezione UMAP in due dimensioni delle Action Units (AUs) tra i due gruppi.

cluster chiaramente distinti di punti tra le due popolazioni. Nel nostro caso, le due popolazioni sono perfettamente sovrapposte, non vi è alcun cluster distinguibile di punti appartenenti a una sola popolazione, questo ci permette di concludere che non vi sono differenze significative tra i valori di AUs tra le due popolazioni, andando a supportare ulteriormente i risultati ottenuti confrontando le espressioni medie, e i risultati ottenuti dai classificatori addestrati.

4.2.5 Riepilogo risultati

In sintesi, l’analisi dei dati di espressione facciale ha riportato risultati contrastanti. Attraverso il confronto dei valori di espressioni medi, massimi e minimi, non sono state trovate differenze significative tra i gruppi (p -value minimo osservato = 0.074) (Tabella 2), e tramite la proiezione su due dimensioni dei dati di AUs, utilizzando la tecnica di riduzione di dimensioni UMAP, non è stata identificata la presenza di cluster distinti (Figura 16), indicando che non vi siano differenze significative tra i valori di AUs durante la visione dei due tipi di video. Attraverso l’addestramento di più classificatori, non è stato possibile identificare con buona precisione quale tipo di video lo spettatore stesse guardando a partire dai soli dati di espressione facciale (Precisione media = 52.0%) (Figura 15). Al contrario, l’analisi delle espressioni discrete⁴ ha mostrato una differenza in termini di tendenza centrale dei valori medi e massimi

⁴Neutrale, sorpresa, tristezza, felicità, paura, disgusto, rabbia.

per partecipante dell'espressione del "Disgusto", con una maggiore predominanza dell'espressione durante la visione dei video sintetici (Tabella 3). Tramite *t*-test delle distribuzioni globali di queste espressioni, sono risultate significativamente diverse, in termini di valore medio, le espressioni della "Rabbia", della "Paura", del "Disgusto" e l'espressione neutrale (Tabella 4, con un valore medio maggiore per le prime tre nei video sintetici, e un valore medio maggiore per l'espressione neutrale nei video reali. Un'analisi qualitativa delle distribuzioni di queste, tramite boxplot, conferma questi risultati trovati, risultando più dilatate e con picchi più alti (Figura 14).

Considerando tutto, i risultati trovati suggeriscono che può esserci una influenza dei video sintetici sulle espressioni facciali, producendo un volto più teso e arricciato, associabile alle espressioni discrete della rabbia, della paura e del disgusto. Questo comportamento può essere la manifestazione di tensione, causata da una comprensione minore dei contenuti proposti, oppure dalla percezione di qualcosa di strano o innaturale nei video sintetici, anche senza riuscire a capire che cosa.

4.3 Dati Fisiologici

In questa sezione affrontiamo l'analisi dei dati fisiologici, raccolti tramite il braccialetto EmbracePlus di Empatica. I dati raccolti sono stati:

- Picchi sistolici: picchi di massimo volume di pressione nel sangue.
- ElectroDermal Activity (EDA): livello di sudorazione della pelle, in μS .

Vediamo come i picchi sistolici sono stati utilizzati per estrarre il segnale di battito cardiaco, e come le distribuzioni dei due segnali di battito cardiaco sono state studiate nei due gruppi: reali e sintetici. In seguito, vediamo come i segnali di EDA sono stati preparati e processati automaticamente, utilizzando il package pyEDA [2], e il confronto dei risultati ottenuti tra i due gruppi.

4.3.1 Picchi sistolici, Battito cardiaco

Per spiegare cos'è un picco sistolico, e come questi sono utilizzati, dobbiamo introdurre brevemente il segnale di Blood Volume Pulse (BVP). Il BVP è un segnale misurato tramite la fotopletismografia (Photoplethysmogram, PPG), la quale tramite un sensore ottico, posto a contatto con la pelle, misura le variazioni di pressione nel sangue. Questi cambiamenti sono direttamente collegati al battito del cuore. È possibile vedere in Figura 17 un esempio di un segnale tipico di BVP. Come è possibile vedere, un buon segnale segue una forma caratteristica, caratterizzata da un alternarsi di un picco massimo, detto "picco sistolico", un secondo picco più basso, e un minimo, detto "picco diastolico". Ogni picco massimo (ovvero ogni picco sistolico) è associato

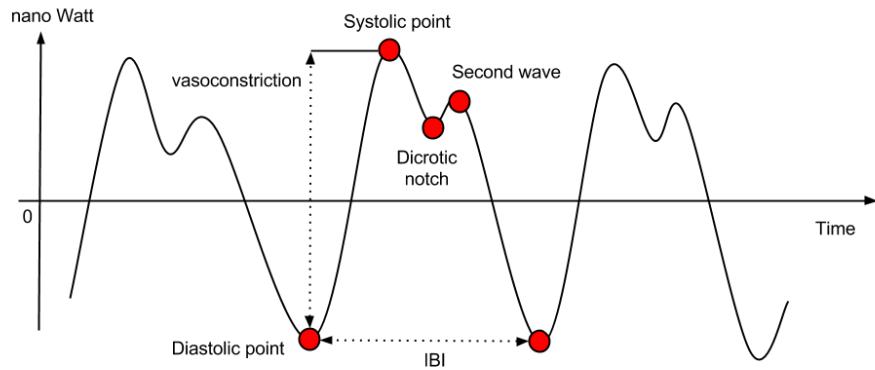


Figura 17: Tipico segnale di BVP. Fonte: <https://support.empatica.com/hc/en-us/articles/360029719792-E4-data-BVP-expected-signal>

a un battito del cuore. Sulla base di questo, l’intervallo di tempo compreso tra due picchi sistolici (o tra due picchi diastolici) è detto Inter-beat-interval (IBI), ovvero il lasso di tempo tra un battito del cuore e l’altro. Avendo a disposizione i timestamp di ogni picco sistolico rilevato, possiamo calcolare la durata di questo intervallo per ogni coppia di battiti e utilizzarlo per stimare localmente il valore del battito cardiaco in quel momento.

Il braccialetto Empatica utilizzato per questa ricerca fornisce sia il segnale grezzo di BVP, che i timestamp dei picchi, estratti utilizzando un algoritmo proprietario, interno al braccialetto. È possibile vedere in Figura 17 una visualizzazione dei dati forniti da Empatica, dove il segnale in blu rappresenta il segnale di BVP, e i puntini rossi rappresentano i picchi sistolici individuati da Empatica. Tramite prove sperimentali, eseguite confrontando l’algoritmo di peak detection nativo di python (utilizzando la libreria `scipy.signal`), e un package gratuito `HeartPy`, specializzato in peak detection per segnali di battito cardiaco, i picchi estratti da Empatica sono risultati essere i più precisi e affidabili, anche in condizione di medio-basso rumore nel segnale.

4.3.2 Calcolo dei BPM

Per estrarre i dati di battito cardiaco dai timestamp dei picchi sistolici è sufficiente implementare un semplice calcolatore di battiti al minuto (Beats Per Minute, BPM). Per ogni coppia consecutiva di picchi sistolici, viene calcolato l’IBI. L’IBI è misurato in secondi per battito. È sufficiente dividere $60 \frac{s}{min}$ (secondi al minuto) per l’IBI

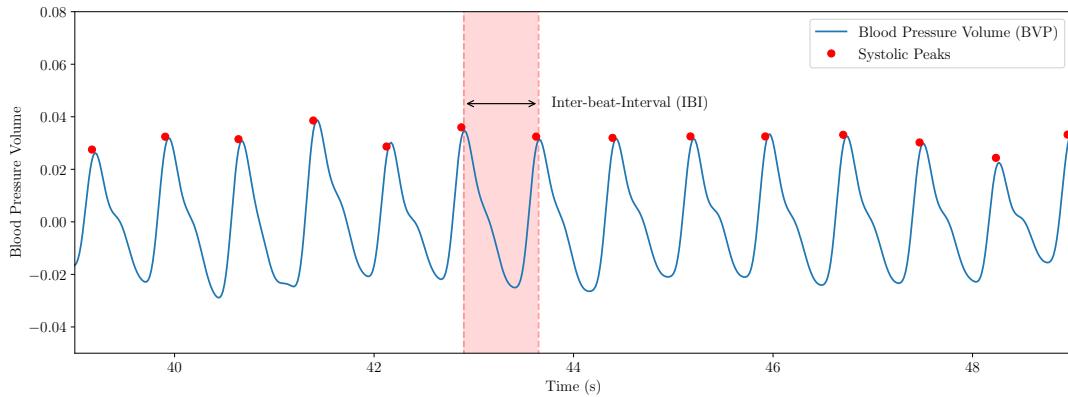


Figura 18: Visualizzazione dei segnali raccolti dal braccialetto Empatica: in blu il segnale di BVP, in rosso i picchi sistolici forniti. L'area evidenziata mostra un esempio di Inter-beat-interval (IBI), valore utilizzato nel calcolo del BPM associato.

appena calcolato, ottenendo una stima del numero battiti al minuto in quell'istante⁵:

$$\text{BPM} = \frac{60}{\text{IBI}} \left[\frac{\frac{s}{\text{min}}}{\frac{s}{\text{Beats}}} = \frac{\text{Beats}}{\text{min}} \right] \quad (5)$$

Ripetendo questa operazione per ogni coppia di picchi sistolici consecutivi presente nei nostri dati andiamo a ricostruire il segnale di battito cardiaco durante l'esperimento.

4.3.3 Analisi del battito cardiaco

Per studiare il comportamento del battito cardiaco durante l'esperimento, sono state confrontate le distribuzioni tra i due gruppi di alcune grandezze, quali la media, la deviazione standard, e la curtosi dei valori di battito cardiaco assunti durante la visione di un video, tramite test di ipotesi appaiati. Per questa parte di analisi statistica è stato utilizzato il tool di analisi automatica autorank [12]. Inoltre, sono stati raccolti tutti i valori di battito cardiaco assunti dai partecipanti, e per i due gruppi, ne è stata confrontata la distribuzione, confrontando qualitativamente i due istogrammi e boxplot, e utilizzando il test di Kolmogorov-Smirnov per confrontare quantitativamente la forma delle due distribuzioni, tramite *p*-value.

⁵Si tratta di una stima in quanto la qualità del risultato dipende dalla qualità dell'identificazione dei due picchi sistolici coinvolti nel calcolo.

Media

È stato calcolato il valore medio di battito cardiaco assunto da ogni partecipante, rispettivamente durante la visione di un video reale e di un video sintetico, ed è stata investigata la presenza di differenze significative tra le due popolazioni. È stato utilizzato come livello di significatività $\alpha=0.050$. Non siamo riusciti a rifiutare l'ipotesi nulla che la popolazione sia normale per tutte le popolazioni (p -value minimo osservato=0.736). Pertanto, assumiamo che tutte le popolazioni siano normali. Poiché abbiamo solo due popolazioni ed entrambe le popolazioni sono normali, utilizziamo il t -test per determinare le differenze tra i valori medi delle popolazioni, e riportiamo il valore medio (M) e la deviazione standard (SD) per ciascuna popolazione. Non siamo riusciti a rifiutare l'ipotesi nulla ($p = 0.092$) del t -test accoppiato, secondo cui i valori medi delle popolazioni “real” ($M=78.589\pm4.783$, $SD=8.790$) e “fake” ($M=79.464\pm4.872$, $SD=8.954$) sono uguali. Pertanto, assumiamo che non vi sia alcuna differenza statisticamente significativa tra i valori medi delle popolazioni.

Deviazione Standard

È stato calcolata la deviazione standard dei valori di battito cardiaco assunti da ogni partecipante, rispettivamente durante la visione di un video reale e di un video sintetico, ed è stata investigata la presenza di differenze significative tra le due popolazioni così create. È stato utilizzato come livello di significatività $\alpha=0.050$. Abbiamo rifiutato l'ipotesi nulla che la popolazione sia normale per la popolazione “real” ($p=0.000$). Pertanto, assumiamo che non tutte le popolazioni siano normali. Poiché abbiamo solo due popolazioni e una di queste non è normale, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. Non siamo riusciti a rifiutare l'ipotesi nulla ($p = 0.507$) del test dei ranghi con segno di Wilcoxon, pertanto, non siamo riusciti a trovare alcuna prova conclusiva di differenze significative tra le popolazioni “fake” ($MD=7.108\pm3.984$, $MAD=2.114$) e “real” ($MD=6.482\pm4.760$, $MAD=0.983$) in termini del valore della deviazione standard dei valori di battito cardiaco.

Curtosi

È stato calcolato l'indice di curtosi associato alle distribuzioni dei valori di battito cardiaco assunti da ogni partecipante, rispettivamente durante la visione di un video reale e di un video sintetico, ed è stata investigata la presenza di differenze significative tra le due popolazioni così create. È stato utilizzato come livello di significatività $\alpha=0.050$. Abbiamo rifiutato l'ipotesi nulla che la popolazione sia normale per le popolazioni “real” ($p=0.001$) e “fake” ($p = 0.005$). Pertanto, assumiamo che non tutte

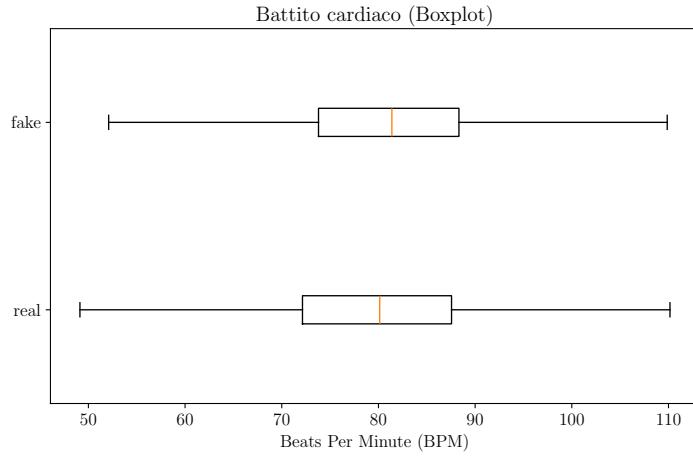
Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
Media	78.589	8.790	79.464	8.790	0.092	inconclusive
Std.	6.482	0.983	7.108	2.114	0.507	inconclusive
Curtosi	2.546	2.131	2.766	2.222	0.392	inconclusive

Tabella 5: Sintesi dei risultati ottenuti analizzando media, deviazione standard e curtosi dei segnali di battito cardiaco acquisiti, utilizzando il package exttautorank. Sono riportate la media (M) e la deviazione standard (SD) per la media, sono riportate la mediana (MD) e la median absolute deviation (MAD) per std. e curtosi.

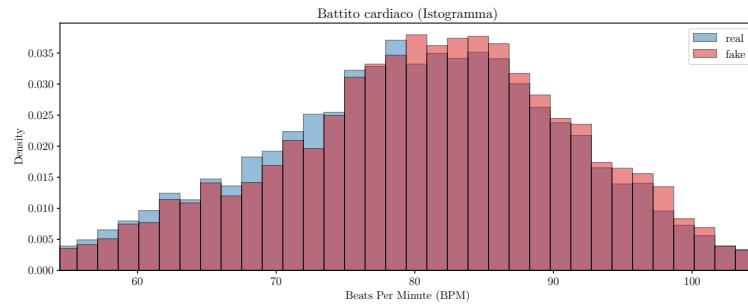
le popolazioni siano normali. Poiché abbiamo solo due popolazioni e una di queste non è normale, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. Non siamo riusciti a rifiutare l’ipotesi nulla ($p = 0.392$) del test dei ranghi con segno di Wilcoxon, pertanto, non siamo riusciti a trovare alcuna prova conclusiva di differenze significative tra le popolazioni “real” ($MD=2.766 \pm 5.071$, $MAD=2.222$) e “fake” ($MD=2.546 \pm 6.627$, $MAD=2.131$), in termini del valore di curtosi dei valori di battito cardiaco.

Test di Kolmogorov-Smirnov

Infine, per confrontare le distribuzioni dei valori di battito cardiaco tra i due gruppi, è stato utilizzato il test di Kolmogorov-Smirnov, un test non parametrico per dati non appaiati, il quale confronta la forma di due distribuzioni, nel nostro caso le distribuzioni di tutti i valori di battito cardiaco assunti durante la visione di un video reale, piuttosto che durante un video sintetico. Il test risponde alla domanda: “quanto è probabile ottenere questi due campioni, se questi fossero campionati dalla stessa distribuzione?” Se il *p*-value è piccolo, concludiamo che i due gruppi sono stati campionati da popolazioni con distribuzioni diverse [13]. Le popolazioni possono differire in media, mediana, variabilità o forma della distribuzione. Il test è stato condotto tra due popolazioni, “real” ($N=10859$) e “fake” ($N=10858$). È stato utilizzato come livello di significatività $\alpha=0.050$. Il test condotto ha indicato la presenza di differenze statisticamente significative ($KS = 0.047$, $p = 8.72 \cdot 10^{-11}$). Rifiutiamo, quindi, l’ipotesi nulla, indicando la presenza di differenze statisticamente significative nella posizione o nella forma delle due distribuzioni. Questo risultato non è indifferente, ma va approfondito, poiché, in casi di dimensione molto elevata del campione, questo tipo di test diventa molto sensibile anche a piccole differenze. Il valore della KS statistic ci informa sulle dimensioni delle differenze trovate.: questa è pari a 0 se le due



(a) Boxplot dei valori di battito cardiaco assunti durante la visione dei due tipi di video.



(b) Istogrammi dei valori di battito cardiaco assunti durante la visione dei due tipi di video.

Figura 19: Visualizzazione grafica della distribuzione dei valori di battito cardiaco tra i due gruppi: (a) boxplot (b) istogramma.

distribuzioni sono identiche, ed è pari a 1 se una distribuzione domina totalmente l'altra. Il valore basso ($KS=0.047$) ci suggerisce che le differenze, sebbene significative, siano piccole. In Figura 19 sono visualizzate le due distribuzioni, tramite boxplot e istogramma. Valutando qualitativamente i grafici, le due distribuzioni hanno una forma molto simile, ma i valori associati ai video fake sono leggermente più alti, con una distribuzione spostata leggermente più sulla destra sull'asse dei valori di battito cardiaco. Nonostante la differenza non significativa tra i valori medi tra i partecipanti, sembrerebbe ci sia una tendenza ad assumere valori di battito cardiaco leggermente più alti durante la visione di video sintetici, sebbene non vi siano abbastanza risultati significativi per poter trarre delle conclusioni definitive a riguardo.

4.3.4 ElectroDermal Activity (EDA)

Per l'analisi dei segnali di ElectroDermal Activity (EDA) è stato utilizzato il package pyEDA [2]. Il package si occupa del pre-processing del segnale di EDA e dell'estrazione di feature associate al segnale. Il package permette di effettuare un'analisi statistica, con l'estrazione di feature statistiche, e un'analisi automatica, con l'estrazione di feature numeriche.

Anche per questo segnale è opportuna una breve spiegazione della sua natura e delle sue caratteristiche, per comprendere al meglio cosa rappresentano le feature e i segnali estratti. L'EDA, noto anche come conduttanza della pelle, o Galvanic Skin Response (GSR), è una misura (in μS) del livello di sudorazione della pelle [4]. Anche se non ce ne accorgiamo, questo valore è costantemente soggetto a piccoli cambiamenti, impercettibili sulla pelle, ma abbastanza grandi da poter essere misurati. La misura è effettuata tramite degli elettrodi posizionati a contatto sulla pelle, ed è stato mostrato come questo segnale è fortemente collegato allo stato psicologico di un individuo, in particolar modo ai livelli di stress e l'attivazione fisica ed emotiva [6]. Diversi fattori possono causare fluttuazioni nel segnale di EDA: naturalmente l'attività motoria, ma anche un forte stimolo emotivo o un segnale di stress.

È possibile vedere in Figura 20 un esempio di un tipico segnale di EDA. Si possono distinguere una componente a bassa frequenza, di lenta crescita e discesa del segnale, e una serie di picchi locali, dovuti a una cresciuta più rapida del segnale. Questo non è un caso. Il segnale può essere scomposto in due componenti: una componente più lenta, detta “tonica”, che comporta il lento crescere e decrescere del livello generale del segnale, e una componente più veloce, detta “fasica”, che fa riferimento a tutti i cambiamenti più repentina del segnale [4]. In particolar modo, in condizioni ottimali la componente fasica è caratterizzata da una serie picchi. Questi picchi rappresentano picchi di attivazione emotiva-sensoriale. A causa di uno stimolo, che può essere un suono, uno stimolo visivo, una reazione emotiva, viene prodotto un picco nel segnale di EDA [4]. Per questi motivi, lo studio di questo segnale si concentra sullo studio dell'analisi del numero dei picchi, per valutare il livello di attivazione emotiva, così come sullo studio del valore medio del segnale e del valore del picco massimo raggiunto tra i gruppi.

In Figura 21 sono mostrate le due componenti di un segnale EDA, estratte mediante il package pyEDA, catturato durante la visione di un video reale. Sono chiaramente visibili i picchi di attività fasica, causati da una visione probabilmente interessata del video mostrato. Andiamo a vedere ora l'analisi che è stata effettuata, a partire dalle feature estratte.

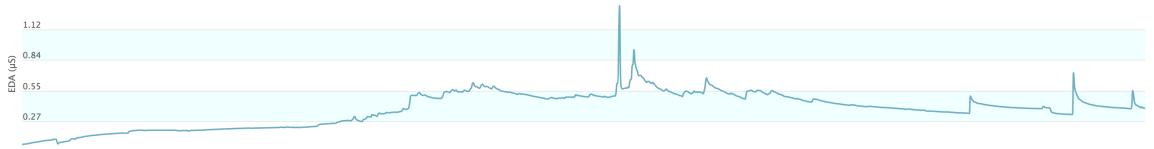


Figura 20: Tipico segnale di EDA, con una componente tonica bassa e dell'attività fasicia (segmento di 35 minuti). Fonte: <https://support.empatica.com/hc/en-us/articles/360030048131-E4-data-EDA-Expected-signal>

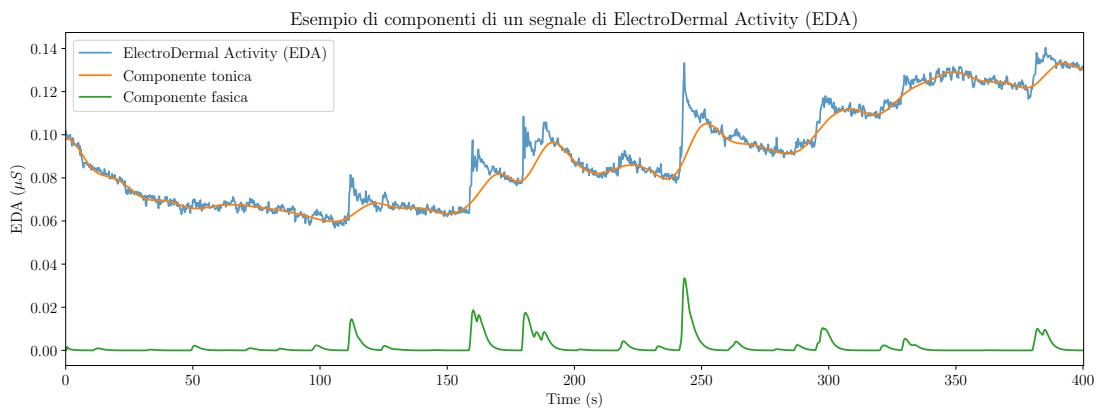


Figura 21: Componenti di un segnale di EDA, estratte con il package pyEDA, catturato durante la visione di un video reale.

4.3.5 Analisi statistica

Le feature estratte dall'analisi statistica sono: il numero di picchi, il valore medio del segnale di EDA e il valore del picco massimo. Il valore di queste feature è stato confrontato tra i due gruppi. Nella Tabella 6 sono riportati i valori medi assunti da ogni feature tra i gruppi. L'analisi statistica è stata condotta per 2 popolazioni ("real" e "fake") con 20 campioni appaiati. È stato utilizzato come livello di significatività $\alpha=0.050$. Per tutte le tre feature, è stato effettuato un test di Shapiro-Wilk per valutare l'ipotesi nulla che le popolazioni non siano normalmente distribuite. Non siamo riusciti a rifiutare l'ipotesi nulla per il numero di picchi (p -value minimo osservato = 0.073), pertanto, abbiamo assunto che questi siano normalmente distribuiti, ed è stato utilizzato un t -test per il confronto, riportando la media (M) e la deviazione standard (SD) per ciascuna popolazione. È stata rifiutata l'ipotesi nulla per i valori di segnale medio e il valore del picco massimo (p -value massimo osservato = 0.000), pertanto, abbiamo assunto che questi dati non siano normalmente distribuiti, ed è stato utilizzato il test dei ranghi con segno di Wilcoxon, riportando la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. Non siamo riusciti

Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
Numero di picchi	89.350	9.207	88.900	9.207	0.893	inconclusive
Segnale medio	0.145	0.102	0.129	0.078	0.464	inconclusive
Picco massimo	0.159	0.112	0.149	0.093	0.522	inconclusive

Tabella 6: Risultati della analisi statistica effettuata, con il package `autorank`, delle feature statistiche dei segnali di ElectroDermal Activity (EDA). Sono riportate media (M) e deviazione standard (SD) per le popolazioni normalmente distribuite (righe evidenziate in grigio), mentre sono riportate mediana (MD) e deviazione mediana assoluta (MAD) per le popolazioni non normalmente distribuite (righe rimanenti).

a rifiutare l’ipotesi nulla per tutti i test condotti, sia nel caso del numero dei picchi ($p = 0.893$), che nel caso del valore medio ($p = 0.464$), che nel caso del picco massimo ($p = 0.522$). In sintesi, In tutti e tre i casi l’analisi è stata inconcludente, ovvero non è stata trovata alcuna prova conclusiva di differenze statisticamente significative tra le feature dei segnali di EDA estratte tra i gruppi. La Tabella 6 riassume i risultati ottenuti. L’analisi è stata effettuata utilizzando il package `autorank`.

4.3.6 Analisi automatica

Il package `pyEDA` permette l’estrazione di feature automatiche, ovvero un array di $k = 32$ feature numeriche $\in \mathbb{R}$. L’estrazione di queste feature ha richiesto la normalizzazione dei segnali nell’intervallo $[0, 1]$ e l’addestramento di un autoencoder. In questo studio, l’autoencoder è stato addestrato per un totale di 100 epoche, con una batch size di 10 sample. Ripetendo l’addestramento, ed effettuando l’analisi con `autorank` più volte, non è stata trovata alcuna differenza statisticamente significativa tra le due popolazioni per tutte le feature estratte (p -value minimo osservato=0.449), rendendo anche questo tipo di analisi sui segnali di EDA inconcludente.

4.3.7 Riepilogo risultati

Facendo un riepilogo dell’analisi sui dati fisiologici, in base ai dati a disposizione, è stata rilevata una tendenza ad assumere valori di battito cardiaco leggermente più alti durante la visione di video lezioni generate tramite IA (Figura 19), sebbene un’analisi dei valori di media, deviazione standard e curtosi non hanno riportato differenze statisticamente significative (Tabella 5).

Per quanto riguarda l’analisi dei segnali di ElectroDermal Activity (EDA), i risultati sono meno interessanti, in quanto, in base ai dati a disposizione, non è stata

Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
HR						
Media	78.589	8.790	79.464	8.790	0.092	inconclusive
Deviazione Standard	6.482	0.983	7.108	2.114	0.507	inconclusive
Curtosi	2.546	2.131	2.766	2.222	0.392	inconclusive
EDA						
Numero di picchi	78.589	8.790	79.464	8.790	0.092	inconclusive
Segnale medio	6.482	0.983	7.108	2.114	0.507	inconclusive
Picco massimo	2.546	2.131	2.766	2.222	0.392	inconclusive

Tabella 7: Risultati dell’analisi di tutti i dati fisiologici. Sono riportate media (M) e deviazione standard (SD) per le popolazioni normalmente distribuite (righe evidenziate in grigio), mentre sono riportate mediana (MD) e deviazione mediana assoluta (MAD) per le popolazioni non normalmente distribuite (righe rimanenti).

trovata sufficiente prova a dimostrare che l’utilizzo di video didattici generati tramite Intelligenza Artificiale (IA) abbia un influenza statisticamente significativa sul livello di attivazione emotiva, misurata tramite analisi del numero di picchi, valore medio del segnale e valore del picco massimo tra i gruppi (Tabella 6).

La Tabella 7 riassume tutti i risultati trovati a seguito dell’analisi dei dati fisiologici.

4.4 Eye-tracking

In questa sezione andiamo ad affrontare l’analisi dei dati di eye-tracking acquisiti. La procedura di analisi è analoga alla procedura seguita per i dati fisiologici: sono state selezionate ed estratte una serie di feature significative, e ne sono state studiate e comparate le distribuzioni. Nel caso dei dati di eye-tracking è stato studiato il comportamento di fissazioni, saccadi e battiti di ciglia. Per le fissazioni, sono state studiate la durata delle fissazioni e il numero di fissazioni al secondo. Per le saccadi, sono state investigate possibili differenze nella durata, l’ampiezza, la velocità massima, la velocità media, e il numero di saccadi per secondo. Per i battito di ciglia, sono state studiate la durata di ogni battito e il numero di battiti di ciglia al secondo. Per ognuna di queste feature è stato confrontato il valore medio, e sono ne state confrontate le distribuzioni tramite K-S test. Per l’analisi dei dati appaiati, quali il confronto dei valori medi, è stato utilizzato il package **autorank**.

	MR	MED	MAD	CI
real	1.684	0.529	0.068	[0.458, 0.835]
fake	1.316	0.620	0.075	[0.527, 0.811]

Tabella 8: Risultati dell’analisi del numero di fissazioni al secondo, eseguita con il package **autorank**.

4.4.1 Fissazioni

Incominciamo l’analisi dei dati di eye-tracking dall’analisi delle fissazioni, confrontando tra i due gruppi la durata delle fissazioni e il numero di fissazioni al secondo. Per il numero di fissazioni al secondo è stato utilizzato il package **autorank**, eseguendo un test tra campioni appaiati. Per la durata delle fissazioni, sono stati confrontati i valori medi tra partecipanti tramite test appaiato, e sono state confrontate le distribuzioni di tutti i valori di durata tra i gruppi, tramite K-S test. Infine, essendo nota la distribuzione della durata delle fissazioni (log normale), sono stati approssimati i parametri di tale distribuzione tra i due gruppi, tramite fitting, e sono stati comparati i valori ottenuti.

Fissazioni per secondo

Incominciamo lo studio del comportamento delle fissazioni dallo studio del numero di fissazioni al secondo per ogni partecipante, durante la visione di un video reale piuttosto che sintetico. L’analisi è stata condotta per 20 campioni appaiati. È stato utilizzato come livello di significatività $\alpha=0.050$. A seguito di un test di normalità di Shapiro-Wilk, abbiamo rifiutato l’ipotesi nulla che la popolazione sia normale sia per la popolazione “real” ($p = 0.001$) che per la popolazione “fake” ($p = 0.000$). Pertanto, assumiamo che non tutte le popolazioni siano normali. Poiché abbiamo solo due popolazioni ed entrambe non sono normali, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. È stata rifiutata l’ipotesi nulla ($p = 0.078$), pertanto, assumiamo che non vi sia alcuna differenza statisticamente significativa tra le mediane delle popolazioni “real” ($MD=0.529\pm0.189$, $MAD=0.068$) e “fake” ($MD=0.620\pm0.142$, $MAD=0.075$). La Tabella 8 riassume i risultati trovati.

Durata delle fissazioni

Per l’analisi della durata delle fissazioni, innanzitutto sono stati confrontati i valori di durata media tra i gruppi.

Media

L'analisi dei valori medi di durata delle fissazioni è stata condotta per 20 campioni appaiati. È stato utilizzato come livello di significatività $\alpha=0.050$. Non siamo riusciti a rifiutare l'ipotesi nulla che la popolazione sia normale per entrambe le popolazioni ($p \geq 0.338$), per tanto, assumiamo che entrambe le popolazioni siano normali. Poiché abbiamo solo due popolazioni ed entrambe sono normali, utilizziamo il t -test per determinare le differenze tra i valori medi delle popolazioni, e riportiamo il valore medio (M) e la deviazione standard (SD) per ciascuna popolazione. Non siamo riusciti a rifiutare l'ipotesi nulla ($p = 0.145$) del t -test accoppiato, per cui i valori medi delle popolazioni "real" ($M=0.404\pm0.069$, $SD=0.127$) e "fake" ($M=0.441\pm0.072$, $SD=0.132$) sono uguali. Pertanto, assumiamo che non vi sia alcuna differenza statisticamente significativa tra i valori medi delle popolazioni.

KS Test

In seguito, sono stati raccolti tutti i dati di durata delle fissazioni associati a un video reale, e tutti i dati di durata delle fissazioni associati a un video sintetico, ed è stato eseguito il test di Kolmogorov Smirnov per confrontare la forma delle due distribuzioni. Il test è stato condotto tra due popolazioni, "real" ($N=14181$) e "fake" ($N=13504$). È stato utilizzato come livello di significatività $\alpha=0.050$. Il test condotto ha indicato la presenza di differenze statisticamente significative tra le due distribuzioni ($KS=0.0269$, $p=0.0001$). Il p -value ottenuto è molto basso, ma, anche in questo caso, la KS statistic molto bassa ($KS=0.0249$) ci suggerisce che le differenze, seppur significative, sono piccole. Questo risultato è probabilmente dovuto all'alta numerosità dei campioni, che rendono il test molto sensibile anche alle piccole differenze.

Fitting su log normale

Possiamo approfondire meglio questa questione, eseguendo un fitting dei dati. Sotto l'assunzione che la fissazione sia l'output di un processo decisionale, il cui reaction time è la durata della fissazione, si può assumere che la distribuzione delle durate sia una log normale. Andiamo, per cui, ad eseguire il fitting dei dati delle due popolazioni su tale distribuzione, il quale ci fornisce una stima dei parametri delle distribuzioni associate, che andiamo a confrontare. Sono state ottenute la distribuzione Lognormal($-0.981, 0.639$) per la popolazione "real", e la distribuzione Lognormal($-0.971, 0.679$) per la popolazione "fake", dove i parametri indicati sono la media e la deviazione standard della distribuzione normale associata⁶. Sono mostrate

⁶Ricordiamo che una distribuzione X , con parametri μ e σ , si dice log normale se $\log(X)$ è distribuita come una distribuzione normale $\mathcal{N}(\mu, \sigma)$.

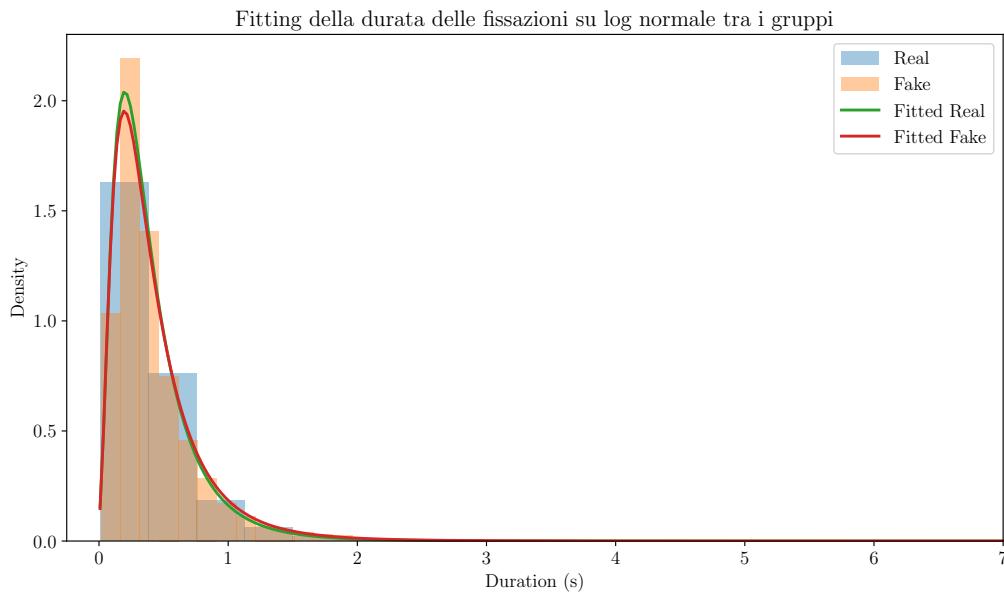


Figura 22: Fitting dei dati di durata delle fissazioni su log normale tra i gruppi.

In Figura 22 le due distribuzioni ottenute tramite fitting. I parametri ottenuti sono abbastanza simili, con una differenza del -0.99% per μ , e una differenza del 6.03% per σ . La differenza tra i parametri suggerisce che le distribuzioni hanno forma molto simile, con la distribuzione associata ai video sintetico leggermente meno concentrata, vista la deviazione standard leggermente più grande. Valutando qualitativamente la figura, la differenza è effettivamente molto piccola, ed è caratterizzata da un picco leggermente più alto nel caso della popolazione “real”, confermando l’ipotesi appena fatta. Per il resto, le due distribuzioni sono estremamente simili, confermando che le differenze trovate tramite KS test, seppur significative, sono molto piccole, ovvero una distribuzione leggermente meno concentrata. Questo risultato è interpretabile come una leggera variabilità in più dei valori di durata delle fissazioni.

4.4.2 Saccadi

Proseguiamo l’analisi dei dati di eye-tracking con lo studio delle saccadi, confrontando tra i due gruppi la durata, l’ampiezza, la velocità massima, la velocità media delle saccadi. Per ognuna di queste feature, sono stati confrontati i valori medi appaiati tra i due gruppi, utilizzando il package `autorank`, e sono state confrontate le distribuzioni di queste feature tra i due gruppi tramite test di Kolmogorov-Smirnov. Infine, è stato confrontato il numero medio di saccadi al secondo durante la visione dei video tra i

due gruppi. Per tutti i test condotti, è stato utilizzato come livello di significatività $\alpha=0.050$.

Durata delle saccadi

Per l'analisi della durata delle saccadi tra i due gruppi sono stati innanzitutto confrontati i valori medi dei valori di durata, calcolati per ogni partecipante tra i gruppi. L'analisi è stata condotta per 2 popolazioni con 20 campioni appaiati. A seguito di un test di normalità di Shapiro-Wilk, è stata rifiutata l'ipotesi nulla che la popolazione sia normale per tutte le popolazioni ($p_{\text{real}} = p_{\text{fake}} = 0.000$), assumiamo, pertanto, che le distribuzioni dei valori medi di durata delle saccadi non siano normalmente distribuite. Dal momento che abbiamo solo due popolazioni, e nessuna è normalmente distribuita, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. In base ai risultati ottenuti, non siamo riusciti a rifiutare l'ipotesi nulla ($p = 0.536$), secondo la quale la popolazione "real" (MD= 0.093 ± 6.528 , MAD=0.040) non è maggiore della popolazione "fake" (MD= 0.092 ± 0.780 , MAD=0.028). Pertanto, assumiamo che non vi siano differenze statisticamente significative tra i valori medi di durata delle saccadi.

Possiamo considerare l'insieme di tutti i valori di durata delle saccadi raccolti durante la visione dei due tipi di video, e confrontarne le distribuzioni. Non è ragionevole assumere che i valori di durata di una saccade siano normalmente distribuiti, non potendo una durata assumere valori negativi, ed essendo i valori molto vicini allo zero. Per questo, non possiamo fare uso di un *t*-test, utilizziamo quindi il test di Kolmogorov-Smirnov, il quale non richiede nessuna assunzione sulla distribuzione dei dati. Il test pone come ipotesi nulla che la distribuzione da cui è stato estratto il primo campione sia la stessa da cui è stato estratto il secondo campione. Il test è stato condotto per due popolazioni, "real" ($N_{\text{real}} = 14171$) e "fake" ($N_{\text{fake}} = 13493$). È stato utilizzato come livello di significatività $\alpha=0.050$. In base ai dati a disposizione, non siamo riusciti a rifiutare l'ipotesi nulla ($KS=0.01$, $p=0.135$). Assumiamo, per cui, che non vi siano differenze statisticamente significative tra la forma delle due distribuzioni dei valori di durata delle saccadi tra i gruppi.

Aampiezza delle saccadi

Per l'analisi dell'ampiezza delle saccadi tra i due gruppi, sono stati innanzitutto confrontati i valori medi di ampiezza, calcolati per ogni partecipante tra i due gruppi.

L'analisi dei valori medi di ampiezza delle saccadi è stata condotta per 2 popolazioni con 20 campioni appaiati. A seguito di un test di normalità di Shapiro-Wilk, è stata rifiutata l'ipotesi nulla di normalità sia per la popolazione "real" ($p = 0.006$) che

per la popolazione “fake” ($p = 0.014$). Pertanto, assumiamo che entrambe le distribuzioni dei valori medi di ampiezza delle saccadi non siano normalmente distribuite. Dal momento che abbiamo solo due popolazioni, e nessuna è normalmente distribuita, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. I risultati sono stati inconcludenti, non indicando la presenza di differenze statisticamente significative ($p = 0.763$) tra i valori medi di ampiezza delle saccadi della popolazione “real” (MD= 1.652 ± 1.477 , MAD= 0.421) e “fake” (MD= 1.719 ± 1.388 , MAD= 0.488).

In seguito, sono state studiate le distribuzioni dei valori di ampiezza delle saccadi registrate durante i due tipi di video. Prima di fare qualsiasi assunzione sulla distribuzione dei dati, effettuiamo un test di Kolmogorov-Smirnov, il quale valuta l’ipotesi nulla che le due popolazioni siano state campionate dalla stessa distribuzione. Il test è stato condotto per due popolazioni, “real” ($N=14171$) e “fake” ($N=13493$). È stato utilizzato come livello di significatività $\alpha=0.050$. È stata rifiutata l’ipotesi nulla ($KS=0.017$, $p=0.038$), suggerendo che, non facendo alcuna assunzione sulla distribuzione dei dati, e in base ai dati raccolti, vi è una differenza statisticamente significativa tra le due distribuzioni dei dati, in termini di forma. Anche in questo caso però, la statistica KS è molto bassa ($KS=0.017$), suggerendo che le differenze trovate da tale test, seppur significative, sono piccole.

Per una analisi più mirata, possiamo fare delle considerazioni sulla distribuzione della durata delle saccadi. Il movimento dello sguardo sullo schermo è paragonabile a un volo di Levy [5], un particolare tipo di random walk in cui le lunghezze dei passi seguono una distribuzione α -stabile [7]. Nel nostro contesto applicativo, la lunghezza dei passi è l’ampiezza delle saccadi. Queste, per cui, seguono, in buona approssimazione, una distribuzione appartenente alla famiglia delle distribuzioni α -stabili. Tra queste, in base a quanto trovato da [5], la distribuzione di Cauchy ricompone bene la distribuzione dell’ampiezza delle saccadi, diversamente ad esempio da una normale. Nelle nostre sperimentazioni, non è stato questo il caso, neanche per la distribuzione di Levy⁷. Per queste ragioni, la scelta è ricaduta sulla distribuzione definita su valori reali positivi che meglio fittava i dati a disposizione: la Inverse Weibull. È stato effettuato, quindi, il fitting dei dati su questa distribuzione, e ne sono state confrontate le distribuzioni confrontando graficamente la loro forma e posizione. In Figura 23, sono riportate e confrontate le distribuzioni ottenute. Come è possibile vedere, le distribuzioni sono molto simili: in questo caso, la differenza più grande è data da un picco leggermente più basso per la distribuzione associata ai video reali, mentre le parti restanti delle due distribuzioni sono approssimativamente sovrapposte. Questo risultato va a confermare la nostra ipotesi, ovvero che le differenze trovate tramite il

⁷La distribuzione di Levy è un’altra distribuzione appartenente alla famiglia delle α -stabili, definita solo su valori reali positivi e a coda lunga.

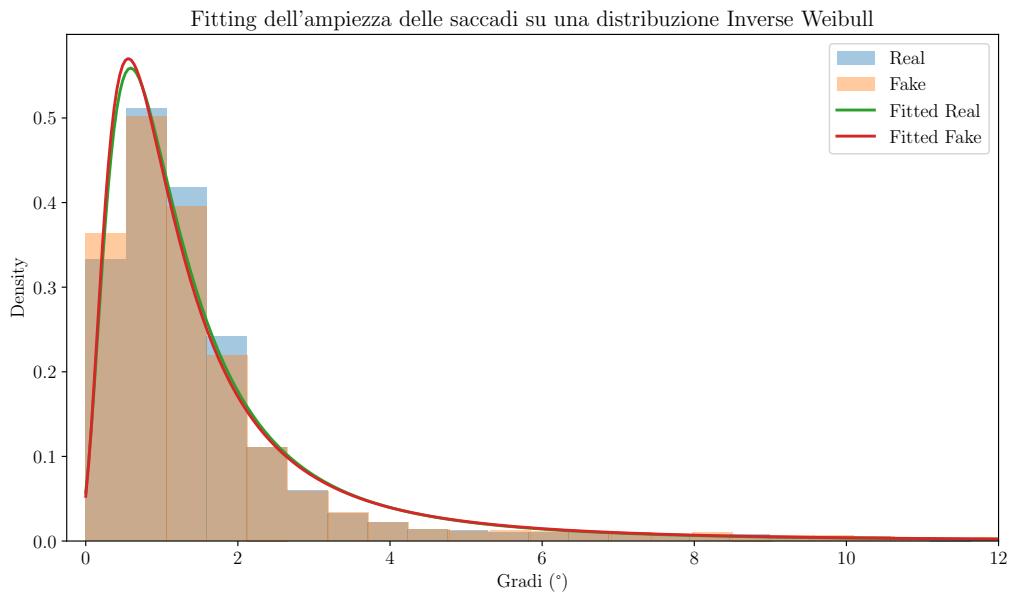


Figura 23: Fitting dei dati di ampiezza delle saccadi su una distribuzione di Cauchy.

test di Kolmogorov-Smirnov, seppure significative, sono molto piccole. La alta sensibilità del test a piccole differenze è probabilmente dovuta alla alta numerosità del campione ($\sim 10^5$ saccadi per popolazione).

Velocità massima delle saccadi

Per l'analisi della velocità massima delle saccadi tra i due gruppi durante la visione dei video, sono state studiate e confrontati valori medi di velocità assunti per partecipante tra i gruppi, ed è stato condotto il testo di Kolmogorov-Smirnov per confrontare le distribuzioni di tutti i dati raccolti tra i due gruppi.

Nel caso della media dei valori di velocità massima delle saccadi, non siamo riusciti a rifiutare l'ipotesi nulla di normalità delle popolazioni ($p \leq 0.058$). Pertanto, assumiamo che entrambe le popolazioni siano normalmente distribuite. Dal momento che abbiamo due popolazioni, ed entrambe sono normalmente distribuite, è stato utilizzato il t -test per campioni appaiati per confrontare le popolazioni. Riportiamo i valori di media (M) e deviazione standard (SD) per ogni popolazione. I risultati sono stati inconcludenti, non indicando la presenza di differenze statisticamente significative ($p = 0.919$) tra i valori medi di velocità massima delle saccadi tra la popolazione “real” ($M=250.965 \pm 53.040$, $SD=97.476$) e la popolazione “fake” ($M=249.148 \pm 47.937$, $SD=88.097$).

In seguito, sono stati considerati gli insiemi di tutti i dati di velocità massima delle saccadi raccolti tra i due gruppi. Non potendo assumere che i dati siano normalmente distribuiti, è stato condotto il test di Kologorov-Smirnov per confrontarne le distribuzioni. Il test è stato condotto per due popolazioni, “real” ($N=14171$) e “fake” ($N=13493$). È stato utilizzato come livello di significatività $\alpha=0.050$. È stata rifiutata l’ipotesi nulla ($KS=0.030$, $p = 5.63 \cdot 10^{-6}$), suggerendo che, e in base ai dati raccolti, vi è una differenza statisticamente significativa tra le due distribuzioni dei dati. Anche in questo caso però, nonostante il p -value estremamente basso, anche la statistica KS è molto bassa ($KS=0.030$), suggerendo che le differenze trovate da tale test, seppur molto significative, sono piccole.

Per l’analisi della velocità media delle saccadi (gradi/s) tra i due gruppi durante la visione dei video, sono state studiate e confrontate la media, la deviazione standard e la curtosí dei valori di velocità media tra i gruppi.

Nel caso della media, non siamo riusciti a rifiutare l’ipotesi nulla di normalità delle popolazioni ($p \geq 0.346$). Pertanto, assumiamo che entrambe le popolazioni siano normalmente distribuite. Dal momento che abbiamo due popolazioni, ed entrambe sono normalmente distribuite, è stato utilizzato il t -test per campioni appaiati per confrontare le popolazioni. Riportiamo i valori di media (M) e deviazione standard (SD) per ogni popolazione. I risultati sono stati inconcludenti, non indicando la presenza di differenze statisticamente significative ($p = 0.933$) tra i valori medi di velocità media delle saccadi tra la popolazione “real” ($M=52.683 \pm 10.926$, $SD=20.079$) e “fake” ($(M=52.856 \pm 8.692$, $SD=15.975$).

In seguito, sono stati considerati gli insiemi di tutti i dati di velocità media delle saccadi raccolti tra i due gruppi. Non potendo assumere che i dati siano normalmente distribuiti, è stato condotto il test di Kologorov-Smirnov per confrontarne le distribuzioni. Il test è stato condotto per due popolazioni, “real” ($N=14171$) e “fake” ($N=13493$). È stato utilizzato come livello di significatività $\alpha=0.050$. Non siamo riusciti a rifiutare l’ipotesi nulla ($KS=0.015$, $p=0.077$), suggerendo che, in base ai dati raccolti, non vi sono differenze statisticamente significative tra le due distribuzioni dei valori di velocità media delle saccadi in termini di forma o posizione.

Saccadi per secondo

L’ultima feature analizzata per lo studio delle saccadi è il numero medio di saccadi per secondo di ogni partecipante durante la visione dei video, tra i due gruppi. A seguito di un test di normalità di Shapiro-Wilk, abbiamo rifiutato l’ipotesi nulla che la popolazione sia normale per entrambe le popolazioni ($p_{\text{real}} = p_{\text{fake}} = 0.000$). Pertanto, assumiamo nessuna delle due popolazioni è normale. Poiché abbiamo solo due popolazioni e nessuna è normale, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD)

Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
# Saccadi/s	0.533	0.076	0.617	0.087	0.508	inconclusive
Durata	0.093	0.040	0.092	0.028	0.536	inconclusive
Aampiezza	1.652	0.421	1.719	0.488	0.763	inconclusive
Velocità massima	250.965	97.476	249.148	97.476	0.919	inconclusive
Velocità media	52.683	20.079	52.856	20.079	0.933	inconclusive

Tabella 9: Sintesi dei risultati ottenuti analizzando i valori medi di durata, ampiezza, velocità massima e velocità minima delle saccadi tra i gruppi, utilizzando il package extttautorank. Sono riportate media (M) e deviazione standard (SD) per le popolazioni normalmente distribuite (righe evidenziate in grigio), mentre sono riportate mediana (MD) e deviazione mediana assoluta (MAD) per le popolazioni non normalmente distribuite (righe rimanenti).

e la deviazione assoluta mediana (MAD) per ciascuna popolazione. Non siamo riusciti a rifiutare l’ipotesi nulla ($p = 0.508$) del test dei ranghi con segno di Wilcoxon, pertanto, non siamo riusciti a trovare alcuna prova conclusiva di differenze statisticamente significative tra le popolazioni due popolazioni “fake” ($MD=0.617\pm0.143$, $MAD=0.087$) e “real” ($MD=0.533\pm0.414$, $MAD=0.076$).

Sintesi dei risultati

Una sintesi de i risultati ottenuti analizzando i valori medi di durata, ampiezza, velocità massima e velocità media delle saccadi, è riportata in Tabella 9. Sono evidenziate in grigio le righe per cui sono riportati i valori di media (M) e deviazione standard (SD). Sono riportati i valori di mediana (MD) e deviazione mediana assoluta (MAD) per le righe rimanenti. Per tutte le feature analizzate, non sono state trovate differenze statisticamente significative tra i valori medi, rendendo questa analisi incocludente. Una sintesi dei risultati ottenuti tramite i test di Kolmogorov-Smirnov, invece, è riportata in Tabella 10. Per tutte le feature analizzate tramite KS test, è stata trovata la presenza di differenze statisticamente significative tra le distribuzioni dei dati tra video reali e video sintetici, con una piccola effect size (KS) per tutte le feature analizzate. La piccola effect size suggerisce che le differenze, seppur significative, sono piccole.

Feature saccadi	<i>KS</i>	<i>p</i> -value	<i>N</i> _{Reale}	<i>N</i> _{Sintetico}
Durata	0.014	0.135	14171	13493
Aampiezza	0.017	0.038	14171	13493
Velocità massima	0.017	0.038	14171	13493
Velocità media	0.015	0.077	14171	13493

Tabella 10: Sintesi dei risultati ottenuti confrontando le distribuzioni di durata, ampiezza, velocità massima e velocità media delle saccadi tra i due gruppi, tramite test di Kolmogorov-Smirnov.

4.4.3 Battiti di ciglia

L’ultima parte di analisi dei dati di eye-tracking è dedicata ai battiti di ciglia. I dati a nostra disposizione sono il timestamp di inizio e fine battito, e la durata del battito di ciglia. Per l’analisi, sono stati analizzate le distribuzioni della durata dei battiti tra le due popolazioni e il numero medio di battiti al secondo eseguiti da ogni partecipante durante la visione dei video proposti.

Numero di battiti al secondo

Iniziamo l’analisi dei battiti di ciglia dal numero medio di battiti al secondo registrati durante la visione dei video per ogni partecipante tra i gruppi. L’analisi statistica è stata condotta per 2 popolazioni (“real” e “fake”) con 20 campioni appaiati. È stato utilizzato come livello di significatività $\alpha=0.050$. A seguito di un test di normalità di Shapiro-Wilk, non siamo riusciti a rifiutare l’ipotesi nulla che la popolazione sia normale sia per le popolazione “real” che per la popolazione “fake” ($p_{\text{real}} = p_{\text{fake}} = 0.000$). Pertanto, assumiamo che nessuna delle due distribuzioni è normalmente distribuita. Dal momento che abbiamo solo due popolazioni, e nessuna è normalmente distribuita, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ciascuna popolazione. Non siamo riusciti a rifiutare l’ipotesi nulla ($p = 0.646$) del test dei ranghi con segno di Wilcoxon, secondo cui la popolazione “fake” ($MD=2.378 \pm 1.869$, $MAD=0.731$) non è maggiore della popolazione “real” ($MD=2.069 \pm 1.904$, $MAD=0.873$). Pertanto, assumiamo che non vi sia alcuna differenza statisticamente significativa tra i valori di mediana del numero di battiti di ciglia al secondo tra i gruppi.

Durata dei battiti di ciglia

Andiamo ora a studiare e confrontare le distribuzioni dei valori di durata dei battiti. Sono stati innanzitutto confrontati i valori di durata media. L’analisi è stata condotta per 2 popolazioni con 20 campioni appaiati. È stata rifiutata l’ipotesi nulla, secondo cui la popolazione è normalmente distribuita, per tutte le popolazioni $p_{\text{real}} = p_{\text{fake}} = 0.000$. Pertanto, assumiamo che nessuna delle due popolazioni sia normalmente distribuita. Dal momento che abbiamo solo due popolazioni, e nessuna è normalmente distribuita, utilizziamo il test dei ranghi con segno di Wilcoxon per determinare le differenze nella tendenza centrale, e riportiamo la mediana (MD) e la deviazione mediana assoluta (MAD) per ogni popolazione. Abbiamo rifiutato l’ipotesi nulla ($p = 0.594$), secondo la quale la popolazione “fake” ($\text{MD} = 0.166 \pm 1.840$, $\text{MAD} = 0.069$) non è maggiore della popolazione “real” ($\text{MD} = 0.149 \pm 19.530$, $\text{MAD} = 0.062$). In luce dei risultati ottenuti, assumiamo che non vi siano differenze statisticamente significative tra i valori di durata dei battiti di ciglia tra le due popolazioni, in termini di mediana.

In seguito, abbiamo considerato l’insieme di tutti i valori di durata registrati durante la visione dei due tipi di video, e abbiamo confrontato le distribuzioni di questi due campioni. Non essendo ragionevole assumere che i valori siano distribuiti normalmente, è stato utilizzato il test di Kolmogorov-Smirnov, che pone come ipotesi nulla che i due campioni siano stati estratti dalla stessa distribuzione. Il test è stato condotto tra due popolazioni, “real” e “fake” ($N_{\text{Reale}} = 4636$, $N_{\text{Sintetico}} = 4304$). È stato utilizzato come livello di significatività $\alpha = 0.050$. In base ai dati a disposizione, è stata rifiutata l’ipotesi nulla ($KS = 0.030$, $p = 0.033$), indicando la presenza di differenze statisticamente significative tra le distribuzioni dei valori di durata del battito di ciglia tra le due popolazioni. Vista l’alta numerosità del campione, è bene osservare il valore della statistica KS. La statistica assume un valore basso ($KS = 0.030$), per cui le differenze trovate, seppur significative, sono piccole. Questo risultato è probabilmente dovuto all’alta numerosità del campione, che rende questo tipo di test sensibile alle piccole differenze.

4.4.4 Riepilogo risultati

In Tabella 11 è fornita una sintesi dell’analisi dei valori medi di tutte le feature dei dati eye-tracking studiate. Per tutte le feature analizzate, non sono state trovate differenze significative tra i valori medi per partecipante ($p \geq 0.075$). In Tabella 12 è riportata una sintesi di tutti i risultati ottenuti dal confronto delle distribuzioni dei valori delle feature studiate tramite test di Kolmogorov Smirnov. Sono riportate la statistica KS ottenuta e il p -value associato, insieme alle numerosità dei campioni studiati. Sono evidenziate in giallo le righe per cui sono state trovate differenze statisticamente significative. Sono state trovate differenze statisticamente significative tra le distribuzioni della durata delle fissazioni ($KS = 0.027$, $p = 9.01 \cdot 10^{-5}$),

Natura video	Reale		Sintetico		<i>p</i> -value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
Fissazioni						
# Fissazioni/s	0.529	0.068	0.620	0.075	0.078	inconclusive
Durata	0.404	0.127	0.441	0.127	0.145	inconclusive
Saccadi						
# Saccadi/s	0.533	0.076	0.617	0.087	0.508	inconclusive
Durata	0.093	0.040	0.092	0.028	0.536	inconclusive
Aampiezza	1.652	0.421	1.719	0.488	0.763	inconclusive
Velocità massima	250.965	97.476	249.148	97.476	0.919	inconclusive
Velocità media	52.683	20.079	52.856	20.079	0.933	inconclusive
Blinks						
# Blink/s	2.069	0.873	2.378	0.731	0.646	inconclusive
Durata	0.149	0.062	0.166	0.069	0.594	inconclusive

Tabella 11: Sintesi dell’analisi dei valori medi di tutte le feature dei dati di eye-tracking analizzate. Sono riportate media (M) e deviazione standard (SD) per le popolazioni normalmente distribuite (righe evidenziate in grigio), mentre sono riportate mediana (MD) e deviazione mediana assoluta (MAD) per le popolazioni non normalmente distribuite (righe rimanenti).

l’ampiezza delle saccadi ($KS = 0.017$, $p = 0.038$), la velocità massima delle saccadi ($KS = 0.017$, $p = 0.038$), e la durata dei battiti di ciglia ($KS = 0.030$, $p = 0.033$). L’elevata numerosità dei due campioni rende il test estremamente sensibile alle piccole differenze. Osserviamo che, per tutti questi risultati, il valore della statistica KS è molto basso ($KS \leq 0.030$), indicando che le differenze trovate, seppur significative, sono di piccole dimensioni. Tutte le altre distribuzioni studiate non hanno riportato differenze statisticamente significative ($p \geq 0.135$).

4.5 Questionari

Concludiamo il capitolo sull’analisi dei dati acquisiti con l’analisi dei dati relativi ai questionari somministrati. I questionario somministrati sono suddivisi in quattro categorie:

- Esperienza di apprendimento: valutazione dell’apprendimento, valutazione del presentatore, somministrati a seguito di ogni video.
- Video Engagement Scale (VES), valutazione dell’esperienza di coinvolgimento durante la visione, somministrato a seguito di ogni video.
- Apprendimento misurato: tre domande a risposta multipla sugli argomenti appena trattati, somministrate a seguito di ogni video.

		KS	<i>p</i> -value	<i>N</i> _{Reale}	<i>N</i> _{Sintetico}
Fissazioni					
	Durata	0.027	$9.01 \cdot 10^{-5}$	14181	13504
Saccadi					
	Durata	0.014	0.135	14171	13493
	Aampiezza	0.017	0.038	14171	13493
	Velocità massima	0.017	0.038	14171	13493
	Velocità media	0.015	0.077	14171	13493
Blinks					
	Durata	0.030	0.033	4636	4304

Tabella 12: Sintesi dei risultati del test di Kolmogorov Smirnov condotto per tutte le feature dei dati di eye-tracking analizzate. Sono riportati i valori della statistica KS ottenuta e i *p*-value associati, insieme alle numerosità dei due campioni. Sono evidenziate in giallo le righe dove sono stati trovate differenze statisticamente significative ($p < 0.050$).

- PANAS: questionario di autovalutazione dello stato emotivo, somministrato a seguito di ogni video.

I questionari somministrati sono forniti nella Sezione 1.1.5. Per ogni item dei questionari proposti è stato, inoltre, registrato il tempo di risposta richiesto dal partecipante per rispondere alla domanda. Per i questionari di valutazione dell'apprendimento, di valutazione del presentatore e per il PANAS, è stata utilizzata una scala di accordo/disaccordo da 1 (“Per niente”) a 5 (“Moltissimo”). Per il questionario Video Engagement Scale (VES) è stata utilizzata una scala numerica da 1 a 7. Inoltre, sono state somministrate una domanda sul grado di familiarità con i contenuti proposti, e una domanda sul grado di utilità percepito del video appena visionato. Anche per queste, è stata utilizzata una scala da 1 (“Per niente”) a 5 (“Moltissimo”).

4.5.1 Risultati dei questionari

Tutti i questionari sono stati definiti in una scala crescente: risposte di basso valore sono associate a valutazioni negative dei video (ad es. basso interesse, basso coinvolgimento), mentre risposte di valore alto sono associate a valutazioni positive dei video. Questa proprietà ci permette di fare una prima analisi, considerando i valori di risposta medi di ogni partecipante ai vari questionari somministrati, per valutare complessivamente l’esperienza di visione di dei video somministrati. Più esplicitamente, sono stati calcolati i risultati medi di ogni partecipante per i dei questionari: esperienza di apprendimento, VES, domande di comprensione, grado di familiarità

Natura video	Reale		Sintetico		<i>p</i> -value
	Media	Sd.	Media	Sd.	
Esperienza di apprendimento	3.45	0.65	2.93	0.73	0.002
Video Engagement Scale (VES)	4.08	0.79	3.59	0.79	0.001
Apprendimento misurato	0.93	0.14	0.81	0.17	0.083
Grado di familiarità	3.22	0.92	3.39	0.83	0.458
Grado di utilità	3.94	0.78	3.17	0.96	0.010

Tabella 13: Per ogni variabile, è fornita la media dei dati raccolti, insieme alla deviazione standard (Sd.). I dati raccolti si basano su risposte espresse su una scala da 1 a 5 per l’esperienza di apprendimento e il grado di utilità, mentre si basano su risposte espresse su una scala da 1 a 7 per il Video Engagement Scale (VES).

e grado di utilità. Dal momento che si tratta di dati ordinali, è stato utilizzato il test dei ranghi con segno di Wilcoxon per il confronto. L’analisi è stata condotta, per ogni questionario, per 2 popolazioni con 20 campioni appaiati. È stato utilizzato come livello di significatività $\alpha=0.050$. Sono state riportate differenze significative nel caso dei questionari di: esperienza di apprendimento ($p = 0.002$), Video Engagement Scale ($p = 0.001$), e grado di utilità ($p = 0.010$). Più nel dettaglio, è stata riportato una esperienza di apprendimento migliore (17.6%), un’esperienza di coinvolgimento maggiore (+13.6%), ed un grado di utilità percepito della lezione somministrata maggiore (+24.6%) a seguito di lezioni realizzate da presentatori reali. Non sono state trovate, invece, differenze significative in termini di apprendimento medio misurato, così come per il grado di familiarità dei contenuti proposti. In Tabella 13 sono riportati i risultati appena presentati, così come i valori di media e deviazione standard per ogni questionario.

4.5.2 Questionari post-visione

Per un’analisi più approfondita, andiamo a fare un confronto dei i risultati ottenuti per ogni item somministrato. Anche qui, trattandosi di dati ordinali, i valori di risposta sono stati confrontati tramite test dei ranghi con segno di Wilcoxon. È stato utilizzato come livello di significanza $\alpha=0.050$. In Tabella 14 sono riportati i risultati ottenuti. Sono riportati i valori di risposta media per ogni domanda tra i gruppi, e per ogni questionario è riportato il valore medio di tutte le risposte del questionario tra i gruppi.

Nel questionario di “Percezione della qualità dell’apprendimento”, sono state individuate differenze statisticamente significative tra i due gruppi per le domande

Risultati dei questionari post-visione

	Reale	AI	p-value
Percezione della qualità dell'apprendimento			
(0) Quanto hai trovato chiari i contenuti della lezione?	4.11	3.44	0.022
(1) Quanto sono stati facili da comprendere i concetti presentati nella lezione?	3.89	3.72	0.426
(2) Quanto hai trovato ben organizzata la presentazione dei contenuti?	3.94	3.50	0.033
(3) Quanto pensi di aver appreso dai contenuti presentati?	3.44	3.06	0.088
(4) Quanto ti senti sicuro/a di ricordare le informazioni apprese nella lezione?	3.28	2.78	0.034
(5) Quanto ti senti preparato/a ad applicare i concetti appresi?	3.28	2.83	0.088
(6) Quanto ritieni utile la lezione per il tuo apprendimento?	3.44	3.00	0.033
(7) Quanto la lezione ha stimolato il tuo interesse per l'argomento?	3.72	3.17	0.031
(8) Quanto ti senti motivato/a a saperne di più sull'argomento dopo aver visto la lezione?	3.39	3.11	0.260
Media dei risultati	3.61	3.18	0.005
Valutazione del presentatore			
(0) Come descriveresti il modo di presentare del relatore?	3.56	2.56	0.002
(1) Quanto hai trovato efficace il relatore nella trasmissione dei contenuti?	3.44	2.78	0.022
(2) Quanto ti è sembrato chiaro e sicuro il relatore durante la spiegazione?	3.67	3.33	0.167
(3) Quanto ti sei sentito/a in sintonia con il relatore?	2.72	2.28	0.101
(4) Quanto il presentatore ti è sembrato naturale e realistico nella presentazione dei contenuti?	3.06	2.50	0.102
Media dei risultati	3.29	2.69	0.006
Video Engagement Scale (VES)			
(0) Durante la visione ero pienamente concentrato sul video.	4.78	4.44	0.224
(1) Durante la visione era come se fossi presente solo a ciò che il video presentava.	4.61	4.33	0.356
(2) Quando stavo vedendo il video, i miei pensieri erano esclusivamente sul video.	4.78	4.22	0.060
(3) Dopo che il video si è concluso, ho avuto la sensazione di essere tornato nel 'mondo reale'.	4.00	3.94	0.892
(4) Dopo un po' di tempo che continuavo a vedere il video, mi è sembrato di diventare una cosa sola con la persona presente nel video.	2.83	2.94	0.837
(5) Mi sono immedesimato nella persona che parlava nel video.	2.83	2.72	0.798
(6) I contenuti del video sono stati coinvolgenti.	5.11	4.00	0.011
(7) Quando stavo vedendo il video, nella mia mente seguivo solo i suoi contenuti.	4.33	3.72	0.325
(8) Durante la visione del video, ho provato le stesse emozioni che provava la persona presente nel video.	3.28	2.50	0.075
(9) Ho trovato il video ingaggiante.	4.39	3.56	0.108
(10) Ho trovato interessante la persona presente nel video.	4.44	2.83	0.005
(11) Durante la visione del video, ero poco attento a cosa ci fosse o a cosa accadesse attorno a me.	4.89	5.17	0.470
(12) Ho avuto la sensazione pensare alle stesse cose che la persona presente nel video diceva.	4.56	4.00	0.282
(13) Nella mia immaginazione, era come se io fossi la persona che parlava nel video.	2.67	2.28	0.469
(14) Grazie al video, mi sono sentito soddisfatto.	3.72	3.22	0.229
Media dei risultati	4.08	3.59	0.081
Domanda di familiarità			
(0) Quanto eri già familiare o a conoscenza dei contenuti mostrati nel video?	3.22	3.39	0.458
Domanda di utilità			
(0) Quanto ti è sembrato utile e/o informativo questo contenuto?	3.94	3.17	0.010

Tabella 14: Risultati dell'analisi delle risposte ai questionari post-visione somministrati, confrontate tramite test dei ranghi con segno di Wilcoxon.

“Quanto hai trovato chiari i contenuti della lezione?” ($p = 0.026$), “Quanto hai trovato ben organizzata la presentazione dei contenuti?” ($p = 0.033$), “Quanto ti senti sicuro/a di ricordare le informazioni apprese nella lezione?” ($p = 0.034$), “Quanto ritieni utile la lezione per il tuo apprendimento?” ($p = 0.033$) e “Quanto la lezione ha stimolato il tuo interesse per l’argomento?” ($p = 0.031$), indicando una preferenza, per tutte queste domande, per i video reali. Complessivamente, i video con presentatore reale sono stati valutati come più chiari, meglio organizzati, e più efficaci nella ritenzione delle informazioni apprese, la percezione dell’utilità della lezione somministrata, e lo stimolo dell’interesse dello spettatore verso l’argomento presentato, rispetto a un video con presentatore generato tramite Intelligenza Artificiale (AI). Nel questionario “Valutazione del presentatore”, sono state individuate differenze statisticamente significative tra i due gruppi per la domanda “Come descriveresti il modo di presentare del relatore?” ($p = 0.002$) e la domanda “Quanto hai trovato efficace il relatore nella trasmissione dei contenuti?” ($p = 0.022$), valutando il modo di presentare del presentatore reale come significativamente più chiaro ed efficace del modo di presentare di un presentatore sintetico. Per il questionario “Video Engagement Scale (VES)”, sono state individuate differenze statisticamente significative per gli item “I contenuti del video sono stati coinvolgenti.” ($p = 0.011$) e “Ho trovato interessante la persona nel video.” ($p = 0.018$), indicando un grado di maggiore coinvolgimento e interesse verso i contenuti presentati da un presentatore reale. È degno di nota il p -value, seppur non significativo, molto basso dell’item “Quando stavo vedendo il video, i miei pensieri erano esclusivamente sul video.” ($p = 0.060$), indicando una possibile tendenza a un grado di attenzione minore durante la visione di un video con un presentatore sintetico. Infine, sono state trovate differenze significative tra i valori di risposta alla domanda di utilità: “Quanto ti è sembrato utile e/o informativo questo contenuto?” ($p = 0.010$), indicando un grado di maggiore utilità percepita della lezione tenuta da un presentatore reale. Tutti gli altri item somministrati non hanno mostrato presenza di differenze statisticamente significative in base ai test condotti.

Effettuando una media di tutti i valori di risposta forniti a ciascun questionario, sono stati valutati come diversi complessivamente diversi i risultati ottenuti dal questionario “Percezione della qualità dell’apprendimento” ($p = 0.005$), indicando di aver percepito di aver appreso di più a seguito di una lezione tenuta da un presentatore reale, e il questionario “Valutazione del presentatore” ($p = 0.006$), indicando un grado di maggiore preferenza verso il presentatore reale. Non sono state trovate differenze statisticamente significative tra i risultati complessivi del questionario “Video Engagement Scale (VES)” ($p = 0.081$), non potendo affermare che vi sia, complessivamente e totalmente, un grado di coinvolgimento maggiore da parte di un tipo di video rispetto all’altro.

4.5.3 Autovalutazione emotiva (PANAS)

Per quanto riguarda i questionari di autovalutazione emotiva (Positive and Negative Affect Schedule, PANAS), sono stati analizzati i risultati ottenuti confrontando le risposte fornite a seguito di un video reale con le risposte fornite a seguito di un video sintetico. Trattandosi di dati ordinali appaiati, per l'analisi è stato utilizzato il test dei ranghi con segno di Wilcoxon. È stato utilizzato come livello di significatività $\alpha=0.050$. Per ogni stato emotivo, l'analisi è stata condotta per 2 popolazioni con 20 campioni appaiati. In Tabella 15 sono forniti, per ogni stato emotivo, il valore medio di risposta indicato tra i gruppi (da una scala da 1 a 5), insieme ai risultati dei test condotti, in termini di p -value. Inoltre, sono stati calcolati i punteggi positivi e negativi medi ottenuti, calcolati sommando i risultati associati rispettivamente ad emozioni positive ed emozioni negative [9]. Per una lettura più immediata, sono evidenziate in tabella in grigio le righe associate ad emozioni negative.

Per tutte le emozioni riportate nel questionario, non sono state trovate differenze statisticamente significative ($p \geq 0.059$). È interessante osservare, però, come tra tutte le emozioni riportate, il p -value più basso è riferito alla voce “Attento” ($p = 0.059$). In base ai dati a disposizione, il risultato non è statisticamente significativo, ma durante l'acquisizione dei dati ha riportato valori di p sempre più piccoli, indicando che, se mantenuta questa tendenza, con un campione più grande si potrebbe provare statisticamente lo stimolo di uno stato di maggiore attenzione percepita a seguito della visione di un video reale. Risulta curioso, invece, lo stato di “Interesse”, il quale non ha mostrato lo stesso comportamento durante lo svolgimento dell'esperimento ($p = 0.254$). Questo risultato, insieme ai risultati sui questionari post-visione, è interessante, in quanto ci suggerisce come l'utilizzo di un video sintetico potrebbe non alterare tanto il grado di interesse verso l'argomento esposto, quanto più la capacità di mantenere attiva l'attenzione dello spettatore, probabilmente a seguito di una efficacia comunicativa e di cattura dell'attenzione significativamente inferiore dei video sintetici, causate da una inferiore chiarezza, naturalezza ed efficacia dei metodi di presentazione ed esposizione di un presentatore sintetico.

4.5.4 Domande di comprensione

Andiamo a indagare meglio la questione, analizzando i risultati delle domande di comprensione, per poter comprendere se questo minor grado di percepito apprendimento e coinvolgimento dello spettatore durante la visione compromette il grado di effettiva comprensione e apprendimento dei contenuti proposti, misurato tramite domande a risposta multipla. Sono state somministrate tre domande a risposta multipla per video, per un totale di sei domande a risposta multipla, con una risposta corretta tra quattro opzioni per ogni domanda. Per ogni video, le tre domande sono state proposte sempre nello stesso ordine, mentre le quattro opzioni di risposta sono state

Positive And Negative Affect Schedule (PANAS)

	Reale	AI	p-value
Nervoso	1.33	1.28	0.715
Attento	3.39	2.94	0.059
Entusiasta	2.83	2.56	0.117
Orgoglioso	2.00	1.89	0.544
Interessato	3.61	3.28	0.254
Deciso	2.67	2.83	0.346
Vigile	3.11	3.00	0.422
Turbato	1.33	1.33	0.726
Vergognoso	1.06	1.11	0.663
Impaurito	1.11	1.22	0.963
Eccitato	1.94	2.06	0.483
Colpevole	1.06	1.11	0.663
Agitato	1.33	1.44	0.446
Ispirato	2.11	2.06	0.578
Irritabile	1.22	1.33	0.674
Angosciato	1.22	1.39	0.433
Ostile	1.11	1.00	0.663
Spaventato	1.17	1.06	0.663
Attivo	3.00	2.94	0.982
Energetico	2.67	2.61	0.963
Punteggio Positivo	27.33	26.17	0.391
Punteggio Negativo	11.94	12.28	0.651

Tabella 15: Risultati del questionario di autovalutazione psico-emotiva PANAS, somministrato a seguito di entrambi i video visionati, con risposta espressa su scala da 1 a 5. Sono evidenziate in grigio le righe associate ad emozioni negative. Sono riportati i valori medi per ogni stato emotivo, e i risultati del test dei segni con rango di Wilcoxon.

Natura video	Reale		Sintetico		p-value
	Media	Std.	Media	Std.	
% risposte corrette	0.93	0.14	0.81	0.17	0.083

Tabella 16: Rapporto medio di risposte corrette alle domande di comprensione tra i gruppi. È indicato il risultato del confronto effettuato tramite test dei ranghi con segno di Wilcoxon, con relativo *p*-value.

Domande di comprensione	% risposte corrette		
	Reale	Sintetico	p-value
Video 1: Comunicazione non verbale			
(0) Quale aspetto della comunicazione non verbale viene considerato cruciale per regolare l'interazione secondo Ray Birdwhistell?	55.6%	88.9%	0.201
(1) Cosa può comunicare una postura chiusa, come le braccia incrociate?	100.0%	100.0%	1.000
(2) Come può un leader efficace utilizzare la comunicazione non verbale?	100.0%	66.7%	0.141
Media dei risultati	85.2%	85.2%	0.789
Video 2: Stress e burnout			
(0) Qual è il gruppo di lavoratori più a rischio di stress secondo il testo?	100.0%	33.3%	0.017
(1) Qual è una delle principali manifestazioni del burnout?	100.0%	100.0%	1.000
(2) Qual è il costo stimato per i datori di lavoro in Europa legato allo stress, ansia e depressione per ogni lavoratore?	100.0%	100.0%	1.000
Media dei risultati	100.0%	77.8%	0.414

Tabella 17: Percentuale di risposte corrette per domanda di comprensione tra i gruppi.

presentate in ordine casuale per ogni partecipante. È stata calcolato un punteggio per ogni partecipante, espresso in percentuale di risposte corrette, tra i gruppi (Tabella 16), la percentuale di risposte corrette per ogni domanda tra i gruppi e la percentuale di risposte corrette totali per video tra i gruppi (Tabella 17).

I punteggi, ovvero le percentuali di risposte corrette tra i gruppi, sono stati confrontati tra i gruppi tramite test dei ranghi con segno di Wilcoxon. È stato utilizzato come livello di significatività $\alpha=0.050$. Il test condotto non ha mostrato la presenza differenze statisticamente significative ($p = 0.083$), non mostrando differenze nel grado di apprendimento complessivo degli argomenti proposti. Analizzando la percentuale di risposte corrette domanda per domanda, sono state riportate differenze significative per le risposte alla domanda “Qual è il gruppo di lavoratori più a rischio di stress secondo il testo?” ($p = 0.017$), prima domanda del video sullo stress e il burnout, con una percentuale di risposte corrette del 100% per i video reali, e del

33.3% per i video sintetici. Non sono state riportate differenze statisticamente significative per le altre domande somministrate. Analizzando la percentuale di risposte corrette per video, non sono state trovate differenze statisticamente significative, con un valore di risposte medie corrette uguale per il primo video (“Comunicazione non verbale”, $M_{Reale} = M_{Sintetico} = 85.2\%$, $p = 0.789$), e non tanto diversi per il secondo video (“Stress e burnout”, $M_{Reale} = 100\%$, $M_{Sintetico} = 77.8\%$, $p = 0.414$). I risultati ottenuti sono riassunti nella Tabella 16, per l’analisi dei punteggi ottenuti, e la Tabella 17, per l’analisi per domanda e per video.

In sintesi, l’analisi delle domande di comprensione non ha riportato sufficienti differenze statisticamente significative per poter trarre delle conclusioni definitive. In altre parole, in base ai dati a disposizione, non vi sono differenze significative nel grado di apprendimento effettivo tra una lezione video tenuta da un presentatore reale e una lezione video tenuta da un presentatore sintetico, nonostante le differenze in termini di chiarezza espositiva e cattura dell’attenzione dello spettatore.

4.5.5 Tempi di risposta

Terminiamo il capitolo di analisi dei dati con l’analisi dei tempi di risposta ad ogni questionario somministrato. Sono stati studiati i tempi di risposta associati agli item dei questionari:

- Percezione della qualità dell’apprendimento
- Valutazione del presentatore
- Video Engagement Scale (VES)
- Domanda di familiarità, Domanda di utilità, PANAS.

Per ogni item di questi questionari, è stato effettuato un test di normalità di Shapiro-Wilk: nel caso di dati valutati come normali è stato effettuato un t -test, mentre nel caso di dati valutati come non normali è stato effettuato un test dei ranghi con segno di Wilcoxon. L’analisi è stata condotta utilizzando il package `autorank`. È stato utilizzato come livello di significatività $\alpha=0.050$. I risultati ottenuti sono riassunti in Tabella 18. Sono evidenziate le righe associate agli item per cui sono state trovate differenze statisticamente significative, e sono indicate con ‘*’ le righe le quali è stato effettuato un t -test.

Per il questionario “Percezione della qualità dell’apprendimento”, sono state riportate differenze statisticamente significative per i tempi di risposta alla domanda “Quanto hai trovato chiari i contenuti della lezione?” (Reale=5.69 s, AI=7.04 s, $p = 0.007$), riportando dei tempi di risposta maggiori a seguito di un video sintetico, indicando una minore sicurezza nel proprio livello di percezione di chiarezza dei contenuti proposti.

Per il questionario “Valutazione del presentatore”, sono state riportate differenze statisticamente significative per i tempi di risposta alla domanda “Quanto ti è sembrato chiaro e sicuro il relatore durante la spiegazione?” (Reale = 4.53 s, AI = 5.41 s, $p = 0.029$), riportando tempi di risposta maggiori a seguito di un video sintetico. Questi risultati indicano maggiore indecisione sulla valutazione del modo di presentare di un presentatore sintetico, nello specifico sulla chiarezza e sicurezza dell’esposizione del presentatore sintetico.

Infine, sono state trovate differenze significative tra i tempi di risposta per la domanda di utilità (“Quanto ti è sembrato utile e/o informativo questo contenuto?”, Reale = 3.62 s, AI = 5.02 s, $p = 0.001$), indicando maggiore indecisione sul grado di utilità dei contenuti presentati da un presentatore sintetico.

Per i questionari rimanenti, ovvero il VES, il PANAS e la domanda di familiarità non sono state riscontrate differenze statisticamente significative tra i tempi di risposta, non indicando differenze significative nel livello di percezione del proprio coinvolgimento con il video, la percezione del proprio stato emotivo, e il livello di familiarità con i contenuti proposti.

4.5.6 Riepilogo risultati

In sintesi, l’analisi dei risultati dei questionari somministrati ha mostrato differenze evidenti in termini di esperienza di visione, coinvolgimento e apprendimento tra video-lezioni presentate da un presentatore reale e un presentatore sintetico. Il questionario “Percezione della qualità dell’apprendimento” ha riportato differenze statisticamente significative tra i gruppi (Reale=3.61, AI=3.18, $p = 0.005$), valutando le lezioni presentate da un presentatore reale come più chiare, ben organizzate, utili, facili da ricordare e stimolanti verso l’argomento presentato. Sono state riportate differenze statisticamente significative tra le valutazioni del presentatore (Reale=3.29, AI=2.69, $p = 0.006$), valutando il presentatore reale come significativamente migliore, nel modo di presentare, in particolar modo nell’efficacia della trasmissione dei contenuti proposti. I video presentati da presentatori reali sono stati valutati come significativamente più coinvolgenti (Reale=5.11, AI=2.83, $p = 0.011$), con un presentatore significativamente più interessante (Reale=4.44, AI=2.83, $p = 0.005$), e il contenuto significativamente più utile e/o informativo (Reale=3.94, AI=3.17, $p = 0.010$). Non sono state riportate, invece differenze significative tra gli stati emotivi autoriportati a seguito delle lezioni somministrate, e non sono state riportate differenze significative tra i risultati delle domande di comprensione somministrate.

Considerando tutti i risultati ottenuti, i video sintetici sono risultati meno chiari e coinvolgenti, stimolando meno l’attenzione e l’interesse dello spettatore verso gli argomenti proposti, i quali sono stati valutati come meno utili e/o informativi. Nonostante questo, non sono state misurate differenze significative in termini di apprendimento

dei contenuti proposti, non indicando che questa differenza sia abbastanza forte da limitare significativamente l'apprendimento effettivo. È bene far notare come la precisione di questo ultimo risultato dipende fortemente dalla dimensione del campione. Avendo quattro combinazioni possibili di test di comprensione⁸, con un campione di 20 partecipanti i dati a nostra disposizione sulle domande di comprensione sono risultati molto limitati, con 4 campioni per test di comprensione. Un campione di maggiori dimensioni potrebbe portare alla luce eventuali differenze nel grado di apprendimento effettivo che noi non siamo riusciti a individuare.

⁸Video 1 fake, Video 2 fake, Video 1 real, Video 2 real.

Domanda	Tempi di risposta (s)		
	Reale	Sintetico	p-value
Percezione della qualità dell'apprendimento			
Quanto hai trovato chiari i contenuti della lezione?	5.69 s	7.04 s	0.007
Quanto sono stati facili da comprendere i concetti presentati nella lezione?	5.60 s	5.75 s	0.490
Quanto hai trovato ben organizzata la presentazione dei contenuti?	5.17 s	4.99 s	0.609
*Quanto pensi di aver appreso dai contenuti presentati?	6.03 s	5.84 s	0.740
*Quanto ti senti sicuro/a di ricordare le informazioni apprese nella lezione?	5.98 s	5.31 s	0.387
Quanto ti senti preparato/a ad applicare i concetti appresi?	6.10 s	6.29 s	0.372
Quanto ritieni utile la lezione per il tuo apprendimento?	6.45 s	6.07 s	0.316
Quanto la lezione ha stimolato il tuo interesse per l'argomento?	5.22 s	5.38 s	0.510
*Quanto ti senti motivato/a a saperne di più sull'argomento dopo aver visto la lezione?	5.38 s	5.39 s	0.987
Valutazione del presentatore			
Come descriveresti il modo di presentare del relatore?	7.54 s	10.18 s	0.072
Quanto hai trovato efficace il relatore nella trasmissione dei contenuti?	6.19 s	6.12 s	0.470
Quanto ti è sembrato chiaro e sicuro il relatore durante la spiegazione?	4.53 s	5.41 s	0.029
Quanto ti sei sentito/a in sintonia con il relatore?	6.28 s	5.68 s	0.281
Quanto il presentatore ti è sembrato naturale e realistico nella presentazione dei contenuti?	7.36 s	7.07 s	0.264
Video Engagement Scale (VES)			
Durante la visione ero pienamente concentrato sul video.	7.06 s	7.58 s	0.647
*Durante la visione era come se fossi presente solo a ciò che il video presentava.	8.85 s	7.60 s	0.318
Quando stavo vedendo il video, i miei pensieri erano esclusivamente sul video.	6.18 s	6.07 s	0.391
Dopo che il video si è concluso, ho avuto la sensazione di essere tornato nel 'mondo reale'.	7.53 s	7.82 s	0.530
*Dopo un po' di tempo che continuavo a vedere il video, mi è sembrato di diventare una cosa sola con la persona presente nel video.	8.07 s	8.78 s	0.531
Mi sono immedesimato nella persona che parlava nel video.	5.35 s	5.16 s	0.628
*I contenuti del video sono stati coinvolgenti.	4.43 s	4.24 s	0.591
Quando stavo vedendo il video, nella mia mente seguivo solo i suoi contenuti.	6.87 s	6.72 s	0.391
Durante la visione del video, ho provato le stesse emozioni che provava la persona presente nel video.	7.54 s	7.14 s	0.530
Ho trovato il video ingaggiante.	4.46 s	4.56 s	0.550
Ho trovato interessante la persona presente nel video.	4.45 s	5.28 s	0.161
Durante la visione del video, ero poco attento a cosa ci fosse o a cosa accadesse attorno a me.	7.13 s	9.11 s	0.088
Ho avuto la sensazione pensare alle stesse cose che la persona presente nel video diceva.	8.24 s	7.92 s	0.353
Nella mia immaginazione, era come se io fossi la persona che parlava nel video.	6.41 s	6.21 s	0.450
*Grazie al video, mi sono sentito soddisfatto.	4.62 s	5.27 s	0.181
Altri			
Quanto eri già familiare o a conoscenza dei contenuti mostrati nel video?	5.05 s	5.22 s	0.874
Quanto ti è sembrato utile e/o informativo questo contenuto?	3.62 s	5.02 s	0.001
PANAS	66.53 s	59.93 s	0.334

Tabella 18: Sintesi dei tempi di risposta media, in secondi, per ogni item dei questionari somministrati, tra i gruppi. Sono riportati i p-value associati. Sono indicati con (*) i dati per i quali è stato effettuato un t-test, poiché valutati come normali dal test di normalità Shapiro-Wilk. È stato effettuato un test dei ranghi con segno di Wilcoxon per i dati restanti.

Capitolo 5

Conclusioni

Questa tesi si è occupata dell'analisi di video lezioni presentate da un presentatore generato tramite Intelligenza Artificiale (AI). Più precisamente, lo scopo del progetto è stato il confronto degli effetti prodotti da video lezioni con presentatori generati tramite AI e presentatori reali su degli osservatori, monitorando segnali fisiologici, comportamento e livello di apprendimento raggiunto. In particolare sono stati valutati il battito cardiaco e la sudorazione della pelle, il movimento dello sguardo sullo schermo, le espressioni facciali assunte durante la visione, e l'esperienza di visione e di apprendimento, valutate attraverso la somministrazione di questionari di valutazione e domande di comprensione sui contenuti proposti. Il progetto ha richiesto lo sviluppo di un'interfaccia ad-hoc per l'acquisizione dei dati, la selezione di un campione pilota, con conseguenti sessione di acquisizione dei dati. L'esperimento è stato strutturato con un disegno *within-subject*: ogni partecipante ha preso visione di due video-lezioni, una con un presentatore reale, una con un presentatore generato tramite IA, in ordine controbilanciato tra tutto il campione, calcolato in base al numero del partecipante. Durante la visione sono stati acquisiti i segnali indicati, e a seguito di ogni video sono stati somministrati i questionari di valutazione e comprensione dei contenuti. Il progetto è stato svolto in collaborazione con il dipartimento di Psicologia dell'Università Cattolica del Sacro Cuore. I testi delle lezioni proposte sono stati prodotti e redatti da Matteo Scarinzi, tesista in Psicologia del Prof. Andrea Gaggioli, e la scelta, redazione e messa appunto dei questionari somministrati è stata svolta dal Prof. Andrea Gaggioli e il Prof. Maurizio Mauri. I video reali sono stati prodotti autonomamente, in collaborazione con due dottorandi del dipartimento di Psicologia della Cattolica, Michele Paleologo e Marta Pizzolante, i quali hanno prestato la loro immagine per i video, e i video sintetici sono stati prodotti utilizzando la piattaforma HeyGen¹.

È stato raccolto un campione di 20 partecipanti. In conclusione, come documentato in Tabella 7, non sono state riscontrare differenze significative nel comportamento

¹www.heygen.com

dei segnali fisiologici. Per il battito cardiaco, sono state confrontate le distribuzioni delle feature di media, deviazione standard e curtosi per partecipante tra i gruppi, e non sono risultate differenze statisticamente significative (Tabella 5). Questo risultato conferma le nostre previsioni, in quanto abbiamo valutato come improbabile che la visione di un video possa alterare significativamente il battito cardiaco. Nel caso della sudorazione della pelle (ElectroDermal Activity, EDA) sono state estratte le seguenti feature: numero di picchi, il segnale medio, e il valore del picco massimo, per partecipante. Queste feature sono associate al livello di attivazione emotiva del partecipante [10], in quanto stimoli esterni possono stimolare la produzione di picchi e la crescita del segnale. Anche queste feature non hanno mostrato differenze significative tra i gruppi, suggerendo che, in base ai dati a disposizione, le video lezioni con presentatore sintetico non hanno dimostrato capacità inferiori di stimolare l'attivazione emotiva dei partecipanti coinvolti rispetto a una video lezione con presentatore reale. Questo risultato era parzialmente previsto, in quanto l'unica differenza visiva tra i video è la natura del presentatore, e in particolare il suo movimento e il suono della sua voce. L'aspetto dei presentatori, così come lo sfondo, sono mantenuti uguali tra i video, per cui non vi sono grandi differenze a livello di stimoli visivi. Detto questo, il segnale di EDA può essere anche stimolato da una reazione emotiva e dagli stimoli auditivi [10], per cui è interessante come anche queste non abbiano comportato delle differenze, riuscendo, in base ai dati raccolti, a replicare fedelmente il livello di attivazione emotiva stimolato dall'esposizione a un video raffigurante una persona reale.

L'analisi delle espressioni facciali ha riportato una presenza significativamente maggiore dell'espressione discreta del disgusto durante la visione dei video sintetici (Tabella 3), in termini di valori medi e massimi. Parliamo di espressioni facciali, quindi questo risultato non intende indicare che sia stata provata maggiormente l'emozione del disgusto, ma la variazione nell'espressione facciale, più tesa e arricciata, potrebbe essere causata da una reazione inconscia, causata dal riconoscimento di qualcosa di strano o innaturale nel video che si sta guardando, anche se non riesce consciamente a capire che cosa. Tutte le altre analisi condotte sulle espressioni facciali non hanno riportato differenze significative, in particolare, l'addestramento di un classificatore tra video reali e sintetici a partire dai soli dati di espressione facciale ha riportato esiti inconcludenti (Figura 15) con una precisione media del 52.0%, rendendo efficace identificare quale tipo di video un partecipante stesse guardando, a partire solo dai dati di espressione facciale, tanto quanto lanciare una moneta.

Il movimento dello sguardo è stato studiato confrontando le distribuzioni di alcune feature estratte, come la durata delle fissazioni, l'ampiezza delle saccadi, e la durata dei battiti di ciglia. Sono state valutate come significativamente diverse le distribuzioni, tramite KS test, (Tabella 12) della durata delle fissazioni, l'ampiezza delle saccadi, la velocità massima delle saccadi e la durata dei battiti di ciglia, suggerendo vi sia una differenza significativa, seppure piccola ($KS < 0.030$), nella forma

delle distribuzioni di queste feature tra i gruppi. Confrontando i valori medi di queste feature tra i gruppi, non sono state riscontrate differenze significative (Tabella 11). Questo risultato è un po' inaspettato, in quanto ci aspettavamo di trovare differenze più evidenti, causate da possibili artefatti o elementi di distrazione, potenzialmente dovuti alla natura sintetica dei video, ad esempio un'imprecisione nel labiale, nel movimento di una parte del corpo. Invece, i risultati trovati suggeriscono che non vi sono grandi elementi di distrazione.

Per quanto riguarda i questionari, i video con presentatore generato tramite IA sono stati valutati come significativamente meno chiari, meno stimolanti, e meno coinvolgenti (Tabella 14), riportando un grado di cattura e mantenimento dell'attenzione peggiore rispetto alle controparti con presentatore reale. Nonostante questo, le domande di comprensione somministrate non hanno mostrato differenze significative tra i due gruppi (Tabella 17), suggerendo che le differenze osservate nei questionari di valutazione non sono tali da risultare un ostacolo concreto per l'apprendimento.

La Tabella 19 fornisce un riepilogo di tutte le analisi condotte tra valori appaiati per partecipante tra i gruppi. Come è possibile vedere in Tabella, le ricerche di differenze tra le feature sono risultate tutte inconcludenti, tranne nel caso dei questionari, e nel caso dell'espressione discreta del disgusto.

Dal momento che non sono state trovate differenze in termini di coinvolgimento emotivo, non sono stati rilevati grandi elementi di distrazione, e i questionari hanno riportato che, nonostante a fronte di una minore chiarezza, coinvolgimento, ed efficacia del presentatore, non sono state rilevate differenze significative in termini di apprendimento, concludiamo che, in base alla tecnologia attualmente disponibile, l'integrazione di video lezioni con presentatori generati tramite IA comporta ancora un compromesso: in base ai nostri risultati, non vi sono grandi differenze in termini di apprendimento, ma l'attenzione, il coinvolgimento e l'interesse degli spettatori verso i contenuti e gli argomenti trattati è significativamente meno stimolato, comportando potenzialmente un calo della performance dello studente nel lungo termine. Detto questo, la generazione dei video è risultata estremamente semplice, riducendo di molto i tempi di produzione, per un costo dei servizi utilizzati ragionevole. Per queste ragioni, la scelta dell'utilizzo e dell'integrazione di queste tecnologie all'interno di un percorso di formazione è consigliabile, ma cercando di garantire sempre un grado di umanità nei contenuti proposti. Facendo qualche esempio, si presenta probabilmente meno adatta per percorsi accademici a lunga durata, dove l'interazione e il rapporto umano con il docente è fondamentale, e il mantenimento dell'attenzione e dell'interesse verso la materia è critico, ma risulterebbe molto conveniente ed efficace per corsi online, o brevi corsi di formazione, ad esempio nelle aziende, dove i costi di produzione di questi corsi sono sempre più importanti.

5.1 Lavori futuri

Questo lavoro è da considerarsi un punto di partenza nella ricerca e nello studio dell'utilizzo di queste nuove tecnologie nel mondo dell'apprendimento. Avendo trattato e analizzato ogni tipo di dato acquisito, le analisi svolte sono state a volte poco specifiche, e si potrebbe entrare nel dettaglio del campo applicativo di ogni tipo di dato analizzato. Ad esempio, nel caso dell'eye-tracker, si potrebbe effettuare una analisi posizionale, andando ad analizzare tramite heat map, per partecipante o accumulata, quali parti dello schermo, o in altre parole, del volto o del corpo del presentatore sono state guardate più a lungo durante la visione dei due video, indagando più direttamente la presenza di possibili fattori di distrazione nei video generati tramite IA. Nel caso dei segnali fisiologici, quali battito cardiaco e sudorazione della pelle, si potrebbero utilizzare tecniche di analisi nello spettro delle frequenze (es. Wavelet Analysis), per l'estrazione e valutazione di feature più complesse, come l'estrazione e l'analisi del ritmo respiratorio dal segnale di battito cardiaco, o un'analisi temporale, analizzando il comportamento di questi segnali durante la visione nel tempo. Si potrebbe estendere l'analisi anche al genere del presentatore, analizzando in maniera incrociata l'influenza del genere del presentatore sugli effetti dell'utilizzo dell'IA. Si potrebbe suddividere l'analisi anche in base ai partecipanti, distinguendo tra studenti e lavoratori, o in base al loro genere, o al livello di educazione, per valutare l'influenza del genere, livello di educazione o della propria occupazione sugli effetti dell'utilizzo di video con presentatori generati con IA. Infine, in generale, si potrebbe ripetere l'esperimento con un campione di dimensioni più elevate. L'utilizzo di un campione più grande potrebbe, innanzitutto, far emergere differenze meno evidenti in un campione di piccole dimensioni, come differenze nel comportamento dei segnali fisiologici, del movimento dello sguardo o delle espressioni facciali. In aggiunta, sarebbe molto importante utilizzare un campione più grande per ottenere risultati più affidabili nel caso dei questionari di comprensione. Visti i quattro possibili questionari somministrati, questi dividono il campione in quattro, richiedendo un campione molto elevato per ottenere dei risultati statisticamente robusti.

Natura video	Reale		Sintetico		p-value	Decisione
	M/MD	SD/MAD	M/MD	SD/MAD		
AUs						
Espress. media	0.278	0.168	0.280	0.168	0.564	inconclusive
Espress. massima	0.882	0.112	0.895	0.112	0.321	inconclusive
Espress. minima	0.025	0.025	0.015	0.015	0.074	inconclusive
Espressioni (%)						
Rabbia	0.923	0.640	0.827	0.543	0.156	inconclusive
Disgusto	0.110	0.089	0.163	0.125	0.007	real < fake
Paura	0.182	0.133	0.156	0.116	0.885	inconclusive
Felicità	0.244	0.127	0.233	0.148	0.551	inconclusive
Tristezza	4.549	4.189	4.850	4.002	0.435	inconclusive
Sorpresa	0.408	0.395	0.678	0.640	0.364	inconclusive
Neutrale	84.535	11.777	86.464	10.294	0.565	inconclusive
HR						
Media	78.589	8.790	79.464	8.790	0.092	inconclusive
Deviazione Standard	6.482	0.983	7.108	2.114	0.507	inconclusive
Curtosi	2.546	2.131	2.766	2.222	0.392	inconclusive
EDA						
Numero di picchi	78.589	8.790	79.464	8.790	0.092	inconclusive
Segnale medio	6.482	0.983	7.108	2.114	0.507	inconclusive
Picco massimo	2.546	2.131	2.766	2.222	0.392	inconclusive
Fissazioni						
# Fissazioni/s	0.529	0.068	0.620	0.075	0.078	inconclusive
Durata	0.404	0.127	0.441	0.127	0.145	inconclusive
Saccadi						
# Saccadi/s	0.533	0.076	0.617	0.087	0.508	inconclusive
Durata	0.093	0.040	0.092	0.028	0.536	inconclusive
Aampiezza	1.652	0.421	1.719	0.488	0.763	inconclusive
Velocità massima	250.965	97.476	249.148	97.476	0.919	inconclusive
Velocità media	52.683	20.079	52.856	20.079	0.933	inconclusive
Blinks						
# Blink/s	2.069	0.873	2.378	0.731	0.646	inconclusive
Durata	0.149	0.062	0.166	0.069	0.594	inconclusive
Questionari						
Esperienza di apprendimento	3.45	0.65	2.93	0.73	0.002	real > fake
Video Engagement Scale (VES)	4.08	0.79	3.59	0.79	0.001	real > fake
Apprendimento misurato	0.93	0.14	0.81	0.17	0.083	inconclusive
Grado di familiarità	3.22	0.92	3.39	0.83	0.458	inconclusive
Grado di utilità	3.94	0.78	3.17	0.96	0.010	real > fake

Tabella 19: Riepilogo di tutte le analisi condotte confrontando i valori per partecipante delle feature indicate tra i gruppi, utilizzando il package **autorank**. Sono riportate media (M) e deviazione standard (SD) per le popolazioni normalmente distribuite (righe evidenziate in grigio), mentre sono riportate mediana (MD) e deviazione mediana assoluta (MAD) per le popolazioni non normalmente distribuite (righe rimanenti).

Bibliografia

- [1] Mike Allen. *The SAGE encyclopedia of communication research methods*. SAGE publications, 2017.
- [2] Seyed Amir Hossein Aqajari, Emad Kasaeyan Naeini, Milad Asgari Mehrabadi, Sina Labbaf, Nikil Dutt, and Amir M Rahmani. pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity. *Procedia Computer Science*, 184:99–106, 2021.
- [3] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022.
- [4] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.
- [5] Dirk Brockmann and Theo Geisel. The ecology of gaze shifts. *Neurocomputing*, 32:643–650, 2000.
- [6] Neil R Carlson. *Physiology of behavior*. Pearson Higher Ed, 2012.
- [7] Alexei V Chechkin, Ralf Metzler, Joseph Klafter, and Vsevolod Yu Gonchar. Introduction to the theory of lévy flights. *Anomalous transport: Foundations and applications*, pages 129–162, 2008.
- [8] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4):781–796, 2023.
- [9] John R Crawford and Julie D Henry. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265, 2004.
- [10] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Unobtrusive assessment of students’ emotional engagement during lectures using electrodermal activity

- sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018.
- [11] Nadeesha M Gunaratne, Claudia Gonzalez Viejo, Thejani M Gunaratne, Damir D Torrico, Hollis Ashman, Frank R Dunshea, and Sigfredo Fuentes. Effects of imagery as visual stimuli on the physiological and emotional responses. *J*, 2(2):206–225, 2019.
 - [12] Steffen Herbold. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020.
 - [13] Joseph Lawson Hodges Jr. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.
 - [14] Janet Shibley Hyde. The gender similarities hypothesis. *American psychologist*, 60(6):581, 2005.
 - [15] Pragma Kar, Samiran Chattopadhyay, and Sandip Chakraborty. Gestatten: Estimation of user’s attention in mobile moocs from eye gaze and gaze gesture tracking. *Proceedings of the ACM on Human-Computer Interaction*, 4(EICS):1–32, 2020.
 - [16] Shengxi Liu, Xiaomei Tao, and Qiong Gui. Research on emotional state in online learning by eye tracking technology. In *Proceedings of the 4th international conference on intelligent information processing*, pages 471–477, 2019.
 - [17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
 - [18] Amy Pitchforth. *Emotional response to auditory and visual stimuli*. Loma Linda University, 2010.
 - [19] Angelica M Tinga, Tycho T De Back, and Max M Louwerse. Non-invasive neurophysiology in learning and training: mechanisms and a swot analysis. *Frontiers in neuroscience*, 14:589, 2020.
 - [20] Leonie NC Visser, Marij A Hillen, Mathilde GE Verdam, Nadine Bol, Hanneke CJM de Haes, and Ellen MA Smets. Assessing engagement while viewing video vignettes; validation of the video engagement scale (ves). *Patient Education and Counseling*, 99(2):227–235, 2016.

- [21] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.
- [22] Zhaoli Zhang, Zhenhua Li, Hai Liu, Taihe Cao, and Sannyuya Liu. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research*, 58(1):63–86, 2020.

Ringraziamenti

*Cara Mamma,
Ce l'ho fatta.*

Anche se non sei qui per vedermi.
mi manchi tanto. Manchi tanto a tutti.
Il tuo bambino si è laureato. E non solo, è felice.

Non posso non sentirmi come di essermi perso una parte che spetterebbe a tutti della vita, ma sto imparando ogni giorno a farne tesoro. Nella buona e nella cattiva sorte. Senza tutto quello che ci hai lasciato, non sarei la persona che sono oggi.

Voglio un mondo di bene alla nonna. Vorrei potesse non morire mai. Vorrei potere io non morire mai, io con lei, però poi so che non ti rivedremmo più. Dio solo sa quanto manchi a lei.

Manchi tanto anche a papà, anche se non lo dà a vedere. È solo merito suo se tutto questo è stato possibile, per aver avuto fiducia in me e aver sacrificato se stesso e i risparmi di una vita per permettermi di vivere e studiare a Milano, cercando di non farmi mancare mai nulla. Come avresti voluto tu. Sarò per sempre in debito per questo, e spero di poter fare sempre il meglio per rendervi entrambi fieri di me, anche quando sarete di nuovo insieme.

Manchi tanto anche ai fratelli grandi, che ogni volta che possono non ci fanno sentire mai soli. E ovviamente anche a Ivan, che si è laureato anche lui, e a cui voglio tanto bene. Io veramente se ho avuto amici a Bari è forse solo merito suo. Devo dire, da allora ho imparato. Anzi, Milano ha giocato molto sporco. Mi ha dato delle persone incredibili, che sono talmente fortunato da poter chiamare famiglia. Spero di poterli non perdere mai.

Per cui non ti preoccupare per me, non sono solo, ci sono i miei amici, e oggi porto addosso i loro vestiti, nel cuore le loro parole, e in bocca il loro sorriso.

*Ciao Mamma
Ti voglio bene*

