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: Prof. Raffaella Lanzarotti

Correlatore: Prof. Andrea Gaggioli

Tesi di:
Federico COSCIA
Matricola: 977772

Anno Accademico 2023-2024

a Celestina

Indice

Introduzione

Capitolo 1

Generazione dei video fake

1.1 Funzionamento

1.2 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
- 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-sync[1], 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.

1.2.1 DupDub

DupDub si classifica come un prodotto "Talking-Photo". A partire da una fotografia di un persona, genera il movimento dei muscoli facciali e delle labbra per simulare il parlato. DupDub trova i suoi punti di forza nell'essere molto semplice, ma è stato valutato come troppo semplice per gli scopi di questa ricerca. La più grande limitazione è data dalla limitatezza dei movimenti, limitandosi appunto a generare solo i movimenti dei muscoli facciali, e a malapena movimenti della testa, rendendo

[1] sincronizzazione tra il movimento delle labbra di un soggetto e il suono delle parole pronunciate.

il risultato finale poco convincente e innaturale. Permette facilmente la creazione di un avatar personalizzato, a partire da un soggetto noto, ma i movimenti sono stati valutati come molto limitati e innaturali, risultando non idoneo per questa ricerca.

1.2.2 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 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 generato. 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.

1.2.3 HeyGen

Sin dal primo sguardo, HeyGen.com 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. Generazione con un avatar personalizzato, creato a partire da un video di riferimento, in grado di clonare l'aspetto e la voce della persona raffigurata

Una volta selezionato l'avatar desiderato, 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. 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 quelli che suonano più naturalmente.

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 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 alla generazione di 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.

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. 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

Si delinea così il tipo di video che siamo in grado di generare: video raffiguranti 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 per la generazione dei video fake
- Generazione di video fake a partire da video registrati con attori, realizzando avatar personalizzati così da avere lo stesso soggetto tra video real e fake

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

1.3 Video generati da video scaricati dal web

1.3.1 Criterio di ricerca

È stata utilizzata per la ricerca dei video la piattaforma YouTube, e il criterio di ricerca dei video reali è semplice: cercare video più simili possibili ai video fake che siamo in grado di generare, così da minimizzare le differenze tra video real e fake. Minimizzando le differenze tra video reali e 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, il criterio di scelta consiste in 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 speaker[2]
- 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

1.3.2 Video trovati

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

[2]questo 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 make a GREAT impression - Presentation Tips"* di Expert Academy
- *"How to start a pitch or presentation"* di Dominic Colenso
- *"How to start a presentation"* di Expert Academy
- *"How to Get Over Your Fear of Public Speaking"* di Expert Academy

I video sono stati scaricati utilizzando il tool open source `yt-dlp` (`https://github.com/yt-dlp/yt-dlp`).

1.3.3 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` (`https://www.ffmpeg.org`), così da favorire un'operazione veloce e priva di operazioni di re-econding 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 dallo speaker, e comparivano solo in sovrapposizione dello sfondo.[3] 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 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

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

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 1.

(a) Fotogramma originale (b) Maschera sovrapposta (c) Fotogramma mascherato

Figura 1: 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.

1.3.4 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 TextToSpeech 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

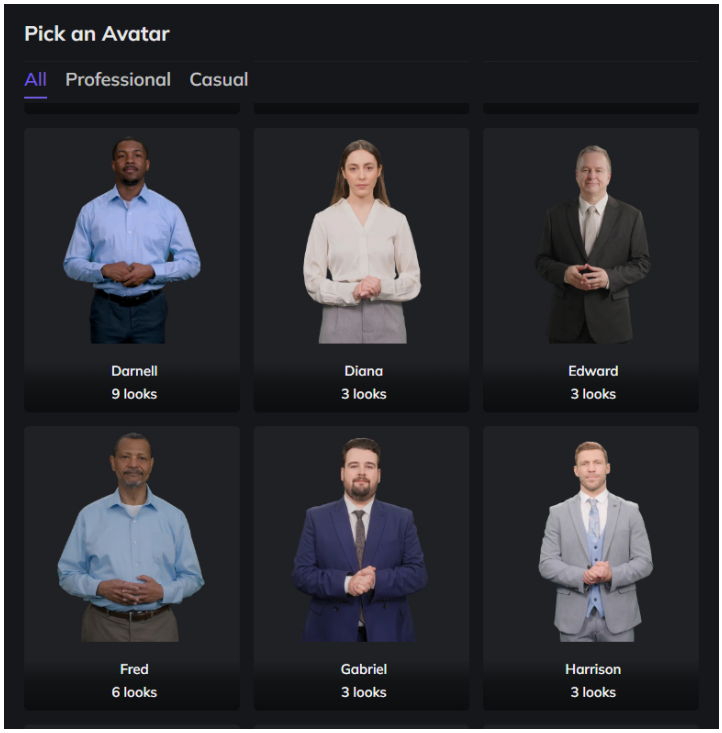

Figura 2: Esempio di schermata di selezione dell'avatar della piattaforma HeyGen

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

1.3.5 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

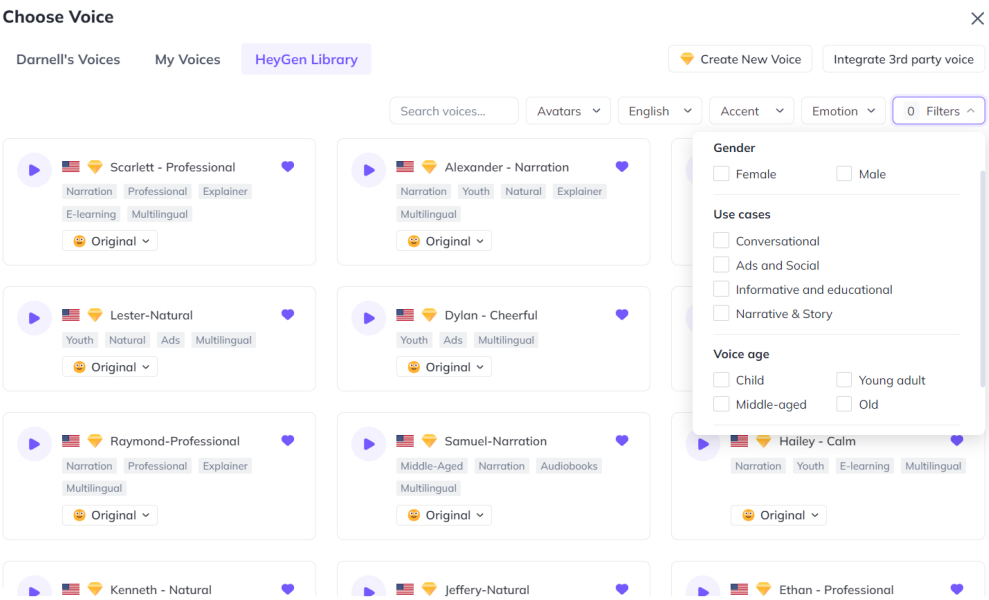

Figura 3: Schermata di selezione della voce sulla piattaforma HeyGen.

figure di persone a mezzo busto o in primo piano, prive di sfondo. È possibile vedere in Figura 2 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.[4]

1.3.6 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, 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

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

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 3, 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.

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.[5]

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.

1.3.7 Scrittura 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 1.3.3. 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

[5]C'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.

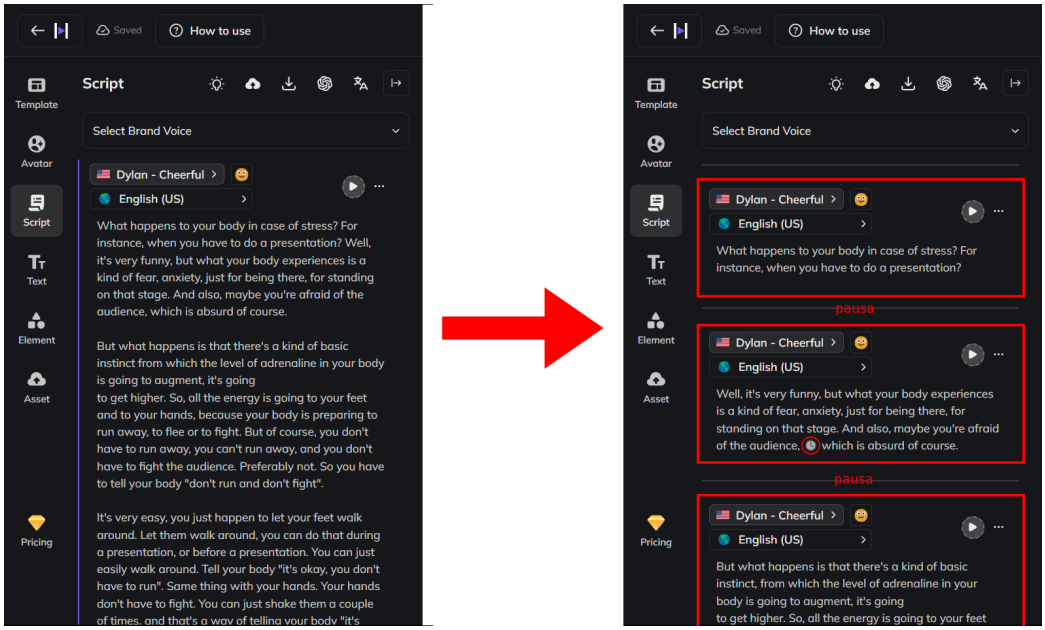

Figura 4: Separazione del testo in paragrafi per introdurre pause naturali.

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 4.

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 1.1. 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.

1.3.8 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.

1.4 Video generati con attori

Capitolo 2

Setting di acquisizione

2.1 Modalità di acquisizione

2.2 Estrazione delle feature

2.2.1 Video

2.2.2 Dati fisiologici

2.2.3 Eye-tracking

Capitolo 3

Protocollo di acquisizione

3.1 Stesura del protocollo

3.2 Sviluppo dell'interfaccia

3.3 Salvataggio dei dati raccolti

Capitolo 4

Analisi dei dati acquisiti

Conclusioni

Bibliografia

Ringraziamenti