POLITECNICO DI TORINO

Corso di Laurea
in Ingegneria Informatica

Tesi di Laurea

Tecnologie di Lipsync Automatico: Un'Analisi di NVIDIA Audio2Face Applicata a Modelli 3D

Relatori
Prof. Andrea Bottino

Candidato
Chiara Sanfilippo

Anno Accademico 2023-2024

Sommario

La crescente richiesta di animazioni facciali realistiche ha stimolato lo sviluppo di tecnologie avanzate per la sincronizzazione labiale nei modelli 3D. Questa tesi esplora l'uso di **NVIDIA Audio2Face** per generare lipsync automatico a partire da tracce audio, analizzando il processo su modelli 3D con o senza blendshape predefiniti.
Attraverso tre casi di studio, viene valutata l'integrazione tra Audio2Face e Unreal Engine con MetaHuman, l'applicazione su modelli dotati di blendshape e l'elaborazione su modelli senza blendshape, importati poi in Autodesk Maya. Sulla base dei risultati ottenuti, la tesi dimostra le potenzialità di Audio2Face nel contesto dell'animazione facciale automatizzata, offrendo soluzioni efficienti per la produzione di contenuti digitali in ambito cinematografico, videoludico e di realtà virtuale.

Ringraziamenti

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione generale al problema

1.1 Animazione di personaggi 3D

Uno degli aspetti più importanti dell'animazione di personaggi 3D è la sincronizzazione dei movimenti facciali con l'audio, nota come lipsync. Questo processo, tradizionalmente gestito manualmente dagli animatori, è diventato sempre più automatizzato grazie alle nuove tecnologie di Intelligenza Artificiale (AI).

1.2 NVIDIA Omniverse Audio2Face

Una delle soluzioni più promettenti in questo ambito è Audio2Face (5), una tecnologia sviluppata da NVIDIA che utilizza l'apprendimento automatico per generare animazioni facciali realistiche a partire da un input audio. Questa tecnologia sfrutta algoritmi di deep learning per analizzare le caratteristiche audio e tradurle in movimenti facciali sincronizzati, consentendo una produzione più efficiente di contenuti animati di alta qualità.

1.3 Obiettivi della tesi

L'obiettivo principale di questa tesi è esplorare le funzionalità di Audio2Face di NVIDIA e valutarne le prestazioni nella realizzazione di lipsync e movimenti facciali su modelli umanoidi.

Il fine ultimo è quello di fornire una comprensione approfondita delle potenzialità offerte da questa tecnologia nel campo dell'animazione di personaggi 3D. L'applicazione Text-to-Speech sviluppata rappresenta un ulteriore contributo per semplificare il processo di generazione di audio da utilizzare come input per Audio2Face.
Questo studio prevede l'analisi delle performance dell'AI applicate alla generazione di blendshape e all'animazione di un modello umanoide, prendendo in input un audio.

Capitolo 2

Introduzione ai task

2.1 Applicazione Text-to-Speech

Per raggiungere questo obiettivo, è stata sviluppata un'applicazione Text-to-Speech basata sulla libreria Streamlit(9). Questa applicazione consente di generare audio partendo da un testo scelto dall'utente, utilizzando la propria voce registrata o un dataset di voci incluso nel progetto. L'applicazione sfrutta due modelli di Hugging Face(4): uno per il Text-to-Speech e uno per la generazione degli embeddings. Il dataset di voci comprende 10 set (6 maschili e 4 femminili) con audio che riproducono frasi in 6 diverse emozioni. Quando viene inserito il testo da generare, l'utente può scegliere il sesso della voce e l'emozione da riprodurre. Tramite un modello di generazione degli embeddings, viene effettuata una text similarity per selezionare l'audio più adatto alla generazione.

2.2 Casi di studio

Questo studio analizza tre casi specifici:

1. Live Streaming dei blendshape da Audio2Face ad Unreal Engine(11) su un modello MetaHuman(10).

2. Generazione di lipsync su un modello che presenta dei blendshape e import su Autodesk Maya(2).

3. Generazione di lipsync su un modello che non presenta dei blendshape e import su Autodesk Maya.

2.2.1 Live Streaming dei blendshape da Audio2Face ad Unreal Engine

In questo caso, viene illustrato come utilizzare la tecnologia Audio2Face per trasferire animazioni facciali basate sull'audio a un modello MetaHuman di Unreal Engine. Una prima fase prevede la realizzazione di un modello MetaHuman partendo da una scansione 3D realizzata tramite fotogrammetria. A supporto di questa fase sono state svolte le seguenti attività preliminari:

- Fotogrammetria e acquisizione del modello 3D.
- Importazione e ottimizzazione del modello in Blender.
- Importazione del modello in Unreal Engine, creazione di una MetaHuman Identity, soluzione dei marker e conversione della mesh in MetaHuman.
- Personalizzazione del MetaHuman tramite MetaHuman Creator.

2.2.2 Generazione di lipsync su un modello che presenta dei blendshape

Questo caso mostra come utilizzare la tecnologia Audio2Face per trasferire animazioni facciali basate sull'audio a un modello umanoide che presenta dei blendshape, ovvero delle deformazioni predefinite della mesh che permettono di animare espressioni facciali.
Le fasi principali di questo processo prevedono:

- La preparazione del modello per l'esportazione da Maya in formato USD e l'importazione di quest'ultimo su Audio2Face.
- Operazioni per l'adattamento della nuova mesh importata e la configurazione della pipeline per il trasferimento delle animazioni.
- Utilizzo di tecniche di mesh fitting che consentono di mappare correttamente i punti di controllo dei blendshape sulla nuova mesh importata.

Infine, viene descritto il processo di conversione dei dati di animazione in un formato compatibile con Autodesk Maya e il loro successivo import nell'ambiente di lavoro 3D per visualizzare e finalizzare le animazioni facciali generate da Audio2Face.

2.2.3 Generazione di lipsync su un modello che non presenta dei blendshape

A differenza del caso precedente, in questo scenario il modello di partenza non possiede blendshape predefiniti. Viene, infatti, illustrato come utilizzare la tecnologia Audio2Face per la generazione automatica di blendshape e, come prima, per trasferire animazioni facciali basate sull'audio a un modello umanoide.
Il processo iniziale è simile a quello del caso precedente, con la preparazione del modello per l'esportazione e il trasferimento della mesh all'interno di Audio2Face. Tuttavia, una volta completato il mesh fitting, viene sfruttata la funzionalità di generazione automatica dei blendshape offerta da Audio2Face.
Questa funzionalità sfrutta algoritmi di apprendimento automatico per analizzare la geometria del modello e generare un set di deformazioni facciali coerenti e realistiche. Vengono esplorate le diverse opzioni di configurazione disponibili per personalizzare il numero e il tipo di blendshape generati, in base alle esigenze specifiche dell'animazione.

Capitolo 3

Lo stato dell'arte

In questo capitolo, ci addentreremo nell'analisi dello stato dell'arte attuale delle tecnologie utilizzate nell'ambito dei metodi di animazione facciale e di sintesi vocale attraverso l'intelligenza artificiale, nota anche come Text-to-Speech. Esploreremo le tecniche più utilizzate in passato, i progressi tecnologici e le applicazioni pratiche di questi strumenti, delineando le sfide e le opportunità che essi offrono nel contesto dell'animazione digitale.

3.1 Metodi di Animazione Facciale

In questa sezione esploreremo le principali tecnologie impiegate nel campo dell'animazione facciale, quali:

- **Animazione 3D tradizionale**: un metodo consolidato che ha rivoluzionato l'industria cinematografica con l'introduzione della grafica computerizzata.
- **Shape keys**: una tecnica utilizzata per creare deformazioni e animazioni su modelli poligonali, ideale per espressioni facciali realistiche.
- **Motion capture**: una tecnologia che cattura i movimenti e le espressioni di attori umani per animazioni facciali altamente realistiche.
- **Animazione con Intelligenza Artificiale**: una tecnologia che sfrutta algoritmi di apprendimento automatico e reti neurali per generare e sincronizzare automaticamente le espressioni facciali.

3.1.1 Animazione 3D tradizionale

L'animazione tradizionale, soprattutto nel contesto dell'animazione 3D, ha segnato una svolta epocale nell'industria cinematografica e dell'intrattenimento. Uno dei momenti più significativi di questa evoluzione è rappresentato dall'uscita di **Toy Story** nel 1995, un film che ha rivoluzionato il settore e ha aperto la strada alla diffusione dell'animazione 3D presso tutti gli studi di animazione.

L'animazione 3D tradizionale è una tecnica che prevede la manipolazione manuale frame by frame, ovvero fotogramma per fotogramma, di modelli 3D digitali da parte degli animatori.
Per garantire che il movimento risultante fosse fluido e realistico era essenziale la definizione manuale di ogni posa con grande precisione. Gli animatori, infatti, dovevano avere un'ottima comprensione dell'anatomia umana e dei principi dell'animazione per creare sequenze fluide e credibili. Si tratta, dunque di un processo lento e laborioso che richiede una grande quantità di lavoro manuale.

3.1.2 Shape Keys

Gli shape keys, conosciuti anche come blendshape o morph target, rappresentano una tecnica fondamentale nell'animazione 3D per generare deformazioni e animazioni di modelli poligonali. Il loro utilizzo ha trasformato radicalmente il modo in cui vengono animate le espressioni facciali e le deformazioni dei modelli, rendendo il processo più efficiente e intuitivo.

L'idea di base degli shape key è creare diverse "forme chiave" dello stesso modello 3D, ognuna rappresentante una specifica deformazione della mesh. Questi vengono poi interpolati e combinati tra loro con pesi diversi durante l'animazione per generare fluidamente un'ampia gamma di espressioni realistiche ed emotive. Questo approccio rese molto più efficiente il lavoro degli animatori facciali rispetto ai metodi tradizionali di manipolazione diretta dei singoli vertici della mesh 3D.

3.1.3 Motion Capture

La motion capture facciale, abbreviata anche in MoCap, è una delle tecniche più avanzate nell'animazione 3D, consentendo di catturare fedelmente i movimenti e le espressioni facciali di attori umani e trasferirli su modelli digitali. Questa tecnologia ha rivoluzionato il settore, aumentando notevolmente il realismo delle animazioni facciali nei film e nei videogiochi.

La motion capture facciale coinvolge l'uso di sensori o marcatori applicati sul viso dell'attore. Sistemi di telecamere o dispositivi appositi rilevano e registrano la posizione di questi marcatori nello spazio durante il movimento. I dati raccolti vengono poi trasferiti su un modello 3D per replicare fedelmente le animazioni facciali. Questo processo permette di catturare movimenti facciali complessi con un elevato grado di precisione.

Vi sono diversi tipi di sistemi di MoCap:

- **Sistemi ottici**: utilizzano telecamere per tracciare marcatori riflettenti o LED posizionati sul viso dell'attore. Questi sistemi sono molto precisi ma richiedono ambienti controllati per evitare interferenze.

- **Sistemi inerziali**: rilevano le rotazioni di sensori inerziali applicati sul corpo. Sono meno suscettibili a interferenze ambientali, ma possono essere meno precisi per i movimenti fini.

- **Sistemi magnetici**: tracciano ricevitori magnetici rispetto a trasmettitori di campo. Offrono buone prestazioni in ambienti chiusi, ma possono essere influenzati da oggetti metallici.

- **Sistemi meccanici**: utilizzano bracci meccanici collegati direttamente agli oggetti da tracciare. Sono estremamente precisi ma limitano i movimenti naturali degli attori.

La MoCap ha quindi la capacità di catturare fedelmente le espressioni umane e rispetto all'animazione manuale frame by frame, è molto più efficiente, riducendo significativamente i tempi di produzione. Tuttavia, si tratta di una tecnologia che può essere costosa, richiedendo attrezzature specializzate e ambienti di registrazione controllati.

3.1.4 Animazione con Intelligenza Artificiale

Negli ultimi anni, l'IA ha rivoluzionato molteplici settori, tra cui l'animazione, grazie allo sviluppo di tecniche avanzate che migliorano notevolmente la qualità e l'efficienza della produzione di animazioni. Un esempio in questo campo è NVIDIA Omniverse Audio2Face.
Audio2Face è una tecnologia avanzata, basata su algoritmi di machine learning e deep learning, che consente di automatizzare il processo di realizzazione di movimenti facciali e sincronizzazioni labiali realistiche a partire da un audio dato in input.

Il funzionamento di Audio2Face si basa su un complesso processo di apprendimento automatico 3.1:

- **Analisi dell'audio**: Audio2Face analizza l'audio di riferimento per estrarre informazioni fondamentali, come il tono della voce, l'intensità delle emozioni trasmesse e il ritmo della parlata.
- **Generazione delle sequenze facciali**: utilizzando reti neurali addestrate su grandi dataset di animazioni facciali e audio correlato, Audio2Face genera sequenze di movimenti facciali coerenti con l'audio di ingresso.
- **Raffinamento manuale**: gli animatori hanno la possibilità di intervenire manualmente per perfezionare le animazioni generate automaticamente da Audio2Face

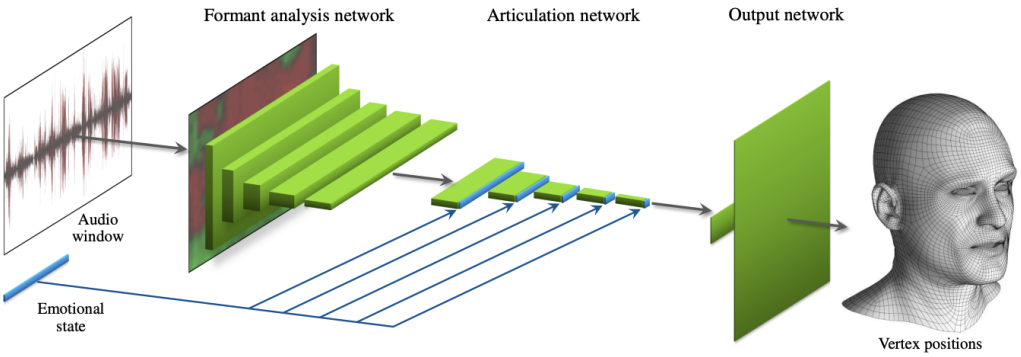


Figura 3.1. Architettura del funzionamento di Audio2Face (6)

3.2 Metodi di Text-to-Speech

In questa sezione esploreremo le principali tecnologie impiegate nel campo del Text-to-Speech, quali:

- **SpeechT5**: un modello Transformer multimodale che consente la conversione tra testo e voce, offrendo una sintesi vocale espressiva e capacità di riconoscimento vocale.

- **Bark**: un modello avanzato di Text-to-Speech in grado di generare voci ricche di emozioni, intonazioni e variazioni contestuali, ideale per narrazioni e dialoghi realistici.

- **XTTS**: una piattaforma open-source flessibile e personalizzabile per la sintesi vocale, che permette la creazione di voci specifiche e multi-lingua per varie applicazioni.

3.2.1 SpeechT5

SpeechT5 rappresenta un'evoluzione significativa nel campo della sintesi vocale e delle applicazioni di elaborazione del linguaggio. Basato sull'architettura Transformer, che ha rivoluzionato numerosi settori del machine learning, questo modello sfrutta reti neurali per convertire il testo in un output vocale naturale e fluido.

Una caratteristica distintiva di SpeechT5 è la sua capacità di eseguire task multimodali, non limitandosi al Text-to-Speech (TTS), ma estendendosi anche allo Speech-to-Text (STT) e alla Voice Conversion. Questa versatilità permette al modello di sintetizzare voce da testo, trascrivere voce in testo e convertire la voce di un parlante in quella di un altro, ampliando notevolmente il suo spettro di applicazioni.

L'addestramento di SpeechT5 avviene su vasti dataset, consentendo al modello di apprendere vari aspetti linguistici come tono, accento ed enfasi. Ciò si traduce in una generazione vocale altamente espressiva e naturale. L'impiego di tecniche di pretraining e fine-tuning permette inoltre di adattare il modello a scenari specifici, migliorando ulteriormente la qualità della sintesi vocale e la capacità di emulare stili o accenti particolari.

3.2.2 Bark

Bark si configura come una delle più recenti innovazioni nel settore del Text-to-Speech (TTS), impiegando modelli di intelligenza artificiale avanzati per produrre una sintesi vocale caratterizzata da elevata naturalezza e ricchezza di sfumature emotive.

A differenza di molte tecnologie TTS tradizionali, che generano voci con una gamma espressiva limitata, Bark è in grado di sintetizzare un output vocale che non solo riproduce accuratamente le parole, ma incorpora anche intonazioni, emozioni e inflessioni contestuali basate sul contenuto del testo. Questa capacità lo rende particolarmente adatto per applicazioni avanzate come la narrazione, il doppiaggio e l'implementazione in chatbot e assistenti virtuali, dove la trasmissione di emozioni umane è cruciale.

3.2.3 XTTS

XTTS si presenta come una soluzione open-source nel campo della sintesi vocale basata su modelli neurali di deep learning.

Una caratteristica distintiva di XTTS è l'elevato grado di personalizzazione offerto. Gli utenti hanno la possibilità di addestrare il modello su dataset vocali specifici, consentendo la creazione di voci uniche e altamente specializzate, in grado di riflettere accenti particolari o stili vocali specifici. Questa flessibilità rende XTTS particolarmente adatto per applicazioni in cui la qualità della voce e la capacità di riprodurre una determinata identità vocale sono fondamentali, come nel caso di videogiochi, produzione di contenuti multimediali o soluzioni di accessibilità.

Capitolo 4

Metodologia

Questo capitolo illustra la metodologia adottata per lo sviluppo di una pipeline che si compone di un sistema integrato di Text-to-Speech e animazione facciale. Il progetto si articola in tre fasi principali:

1. Sviluppo di un'applicazione Text-to-Speech.
2. Generazione di lipsync tramite Audio2Face.
3. Integrazione dell'animazione facciale su Autodesk Maya.

L'approccio metodologico integra tecnologie di deep learning per il Text-to-Speech con tecniche avanzate di computer grafica per l'animazione facciale, utilizzando strumenti all'avanguardia come Hugging Face, NVIDIA Audio2Face e software di modellazione 3D.

4.1 Applicazione Text-to-Speech

L'applicazione Text-to-Speech è stata sviluppata utilizzando il framework Streamlit, che permette di creare rapidamente interfacce web interattive per applicazioni di machine learning.

4.1.1 Modelli e dataset

Il sistema si basa su due modelli principali di Hugging Face:

- facebook/fastspeech2-en-ljspeech che trasforma il testo scritto in parlato.

- sentence-transformers/all-MiniLM-L6-v2 per la conversione del testo in embedding.

Il dataset di voci include 10 set (6 maschili, 4 femminili) con 6 diverse emozioni per ciascuna voce, per un totale di 60 campioni audio.

4.1.2 Processo di selezione dell'audio

La selezione della voce più appropriata avviene tramite un algoritmo di similarità testuale. Questo approccio intelligente consente di abbinare automaticamente il contenuto del testo di input alla voce più adatta, garantendo una maggiore coerenza e naturalezza nell'output audio generato.

Ogni campione audio nel dataset non è solo un file sonoro, ma è accompagnato da una breve descrizione testuale. Queste descrizioni forniscono informazioni sul contenuto, il tono o il contesto del campione audio. Ad esempio:

- "Voce maschile, tono professionale, per presentazioni aziendali"
- "Voce femminile, tono amichevole, per assistenza clienti"
- "Voce giovane, entusiasta, per pubblicità di prodotti sportivi"

Il processo si articola in diverse fasi chiave. Inizialmente, il testo fornito dall'utente viene trasformato in un embedding vettoriale, una rappresentazione numerica che cattura le sfumature semantiche del contenuto. Parallelamente, lo stesso processo di vettorizzazione viene applicato alle descrizioni associate a ciascun campione audio nel dataset.

Una volta ottenuti questi vettori, l'algoritmo procede al calcolo della similarità del coseno tra l'embedding del testo di input e quelli di tutti i campioni audio disponibili. Questa metrica matematica permette di quantificare quanto due testi siano semanticamente vicini nello spazio vettoriale.

Il passaggio finale consiste nella selezione del campione audio la cui descrizione presenta la maggiore similarità del coseno con il testo di input. Questo significa che viene scelta la voce il cui contesto o stile si allinea più strettamente con il contenuto del messaggio da sintetizzare.

Grazie a questo meccanismo, l'applicazione è in grado di selezionare automaticamente la voce più appropriata per ogni input testuale, tenendo conto non solo delle parole utilizzate, ma anche del contesto e del tono generale del messaggio. Ciò contribuisce a produrre un output audio più naturale e contestualmente appropriato, migliorando la qualità complessiva della sintesi vocale.

4.2 Generazione di lipsync tramite Audio2Face

La generazione di lipsync tramite Audio2Face è stata effettuata esplorando i tre casi di studio di seguito riportati:

- Live Streaming dei blendshape da Audio2Face ad Unreal Engine su un modello MetaHuman.
- Generazione di lipsync su un modello che presenta dei blendshape e import su Autodesk Maya.
- Generazione di lipsync su un modello che non presenta dei blendshape e import su Autodesk Maya.

Di seguito verranno analizzati nel dettaglio i tre casi di studio citati in precedenza.

4.2.1 Live Streaming dei blendshape da Audio2Face ad Unreal Engine

Questo caso di studio esplora l'integrazione in tempo reale tra NVIDIA Audio2Face e personaggi MetaHuman in Unreal Engine e si compone di due fasi principali:

1. Creazione di un modello MetaHuman.
2. Live Streaming dei blendshape.

Di seguito verranno esplorate entrambe le fasi.

Creazione di un modello MetaHuman

La creazione di un modello MetaHuman personalizzato implica una serie di processi tecnici che integrano diverse tecnologie di computer vision e grafica 3D. Il workflow si articola in diverse fasi:

1. Acquisizione dei dati tramite fotogrammetria.
2. Ottimizzazione della mesh in Blender.
3. Integrazione in Unreal Engine.
4. Perfezionamento del modello nel MetaHuman Creator.

Acquisizione dei dati tramite fotogrammetria

La fase iniziale del processo di creazione del modello MetaHuman prevede l'utilizzo della fotogrammetria che consente di acquisire un modello 3D di un soggetto attraverso l'analisi e l'interpretazione di immagini fotografiche. Per questa fase è stata impiegata l'applicazione Polycam(8).
Il processo inizia con l'acquisizione di un dataset fotogrammetrico composto da immagini ad alta risoluzione, catturate a 360 gradi intorno al soggetto. Al termine dell'acquisizione, l'applicazione Polycam processa le immagini mediante algoritmi di ricostruzione tridimensionale, generando una mesh poligonale texturizzata del volto del soggetto che viene poi esportata in formato GLTF.

Ottimizzazione della mesh in Blender

La mesh risultante viene importata in Blender(3) per un processo di ottimizzazione. Questa fase include:

1. Applicazione del modificatore Shade Smooth.
2. Rimozione di vertici non necessari.
3. Operazioni di pulizia della mesh, quali:
    - Delete Loose per eliminare vertici ed elementi non connessi.
    - Degenerate Dissolve per dissolvere gli elementi come facce con area nulla o spigoli di lunghezza zero.
    - Merge By Distance per unire i vertici molto vicini tra loro riducendo la ridondanza geometrica.

Questa sequenza di operazioni avrà come risultato una mesh unificata, ottimizzata e priva di artefatti geometrici indesiderati, preparando il modello per le successive fasi di lavorazione.
Il modello viene infine esportato in formato FBX.

Integrazione in Unreal Engine

L'importazione della mesh in Unreal Engine segna l'inizio della fase di trasformazione in MetaHuman. Questa fase comporta:

1. Importazione della mesh FBX nell'ambiente Unreal.
2. Creazione della MetaHuman Identity.
3. Marcatura dei punti di riferimento facciali chiave (landmarks).
4. Risoluzione dell'identità MetaHuman (MetaHuman Identity Solve).

Questo ultimo processo analizza la geometria del modello importato, mappa le caratteristiche facciali sulla struttura standard MetaHuman e le trasforma in un modello MetaHuman che mantiene le caratteristiche distintive del soggetto originale.

Perfezionamento del modello nel MetaHuman Creator

Il processo si conclude con una fase di personalizzazione del modello nel MetaHuman Creator. Qui possiamo regolare sottili dettagli come la texture della pelle, la forma degli occhi, o la precisione delle espressioni facciali, assicurandoci che il nostro MetaHuman non sia tecnicamente accurato e realistico.

Live Streaming dei blendshape

La funzionalità di Live Streaming è resa possibile attraverso un plugin dedicato che funge da ponte tra le due piattaforme. Il processo si articola in tre fasi principali:

1. Installazione del plugin Audio2Face in Unreal Engine.
2. Configurazione del LiveLink per la comunicazione tra Audio2Face e Unreal.

3. Attivazione dello streaming in tempo reale dei blendshape da Audio2Face.

Inizialmente, si procede con la configurazione del plugin Audio2Face all'interno di Unreal Engine, stabilendo i parametri necessari per la comunicazione tra i due software. Successivamente, si instaura un collegamento live, che consente il flusso continuo di dati dall'analisi audio di Audio2Face al motore di rendering di Unreal Engine. Infine, si effettua una mappatura precisa tra i blendshape generati dinamicamente da Audio2Face e i corrispondenti controlli facciali del modello MetaHuman.
Questo approccio permette di animare il personaggio MetaHuman in tempo reale in base all'input audio.

4.2.2 Generazione di lipsync su un modello che presenta dei blendshape

Questo approccio si concentra sull'utilizzo di un modello 3D preesistente con blendshape già definiti. Il workflow si articola in diverse fasi:

1. Preparazione del modello 3D in Maya.

2. Importazione in Audio2Face.

3. Generazione dell'animazione.

4. Esportazione dei dati di animazione.

5. Importazione in Maya.

Preparazione del modello 3D in Maya

È fondamentale che la mesh facciale sia strutturata in modo ottimale, con una topologia pulita e una corretta definizione dei blendshape. Il modello viene quindi esportato in formato USD (Universal Scene Description), un formato file open-source sviluppato da Pixar(7) che offre una rappresentazione flessibile e dettagliata di scene 3D complesse. La scelta del formato USD è strategica, in quanto garantisce una conservazione accurata della geometria e dei dati di animazione durante il trasferimento tra piattaforme diverse.

Importazione in Audio2Face

Una volta importato in Audio2Face, il modello subisce un processo di mesh fitting. Questa fase richiede una mappatura precisa tra i punti di controllo del modello importato e quelli del modello di riferimento interno di Audio2Face. Il processo di mappatura può essere effettuato manualmente o assistito da algoritmi di rilevamento automatico dei punti chiave, con la possibilità di raffinare manualmente i risultati per ottenere la massima precisione.
L'utente deve quindi posizionare manualmente una serie di punti di controllo su aree specifiche del viso del modello importato utilizzando il tool chiamato Character Transfer. Questi punti corrispondono a zone come gli angoli degli occhi, le narici, gli angoli della bocca, il contorno del viso, ecc.
Una volta che i punti di controllo sono posizionati correttamente, Audio2Face calcola come mappare i movimenti facciali generati sul modello importato.

Questo processo assicura che il sistema comprenda correttamente la topologia del viso del modello importato e che l'animazione generata si mappi correttamente sulla topologia specifica del modello dell'utente.

Generazione dell'animazione

La generazione dell'animazione facciale in Audio2Face sfrutta algoritmi avanzati di machine learning per interpretare l'input audio e tradurlo in movimenti facciali realistici.
Nello specifico il processo si svolge nel seguente modo:

1. Analisi dell'input audio:
    - Il sistema analizza l'audio in ingresso utilizzando tecniche di elaborazione del segnale digitale.
    - Questa analisi scompone l'audio in diverse componenti, tra cui:
        (a) Fonemi: le unità sonore di base del linguaggio.
        (b) Prosodia: l'intonazione, il ritmo e l'enfasi nel parlato.
        (c) Intensità: il volume e la forza del suono.
2. Riconoscimento dei fonemi:

- Algoritmi di deep learning, probabilmente basati su reti neurali convoluzionali (CNN) o reti neurali ricorrenti (RNN), identificano i fonemi presenti nell'audio.
- Ogni fonema viene associato a una specifica configurazione della bocca e delle labbra.

3. Analisi prosodica:

   - Modelli di machine learning analizzano la prosodia per catturare:
     (a) Variazioni di tono: che possono indicare emozioni o enfasi.
     (b) Ritmo del parlato: che influenza la velocità e la fluidità dei movimenti facciali.
     (c) Pause e accentuazioni: che possono tradursi in espressioni facciali specifiche.

4. Mappatura sui blendshape:

   - Il sistema utilizza una rete neurale pre-addestrata per mappare le informazioni audio estratte sui blendshapes del modello 3D.

5. Generazione delle curve di animazione:

   - Per ogni blendshape, il sistema genera una curva di animazione nel tempo: che descrivono come l'intensità di ogni blendshape varia durante l'audio.

Il sistema, dunque, non analizza solo i fonemi per il lip-sync, ma anche le sfumature dell'intonazione e dell'enfasi per generare espressioni facciali coerenti. L'output di questo processo è un set di dati di animazione che descrive il movimento di ogni blendshape nel tempo.

Esportazione dei dati di animazione

I dati di animazione ottenuti nella fase precedente vengono esportati in un file formato JSON che contiene informazioni dettagliate sui blendshape seguendo la seguente struttura dei dati:

- "exportFps": frame rate dell'animazione.
- "trackPath": percorso del file audio utilizzato.

- "numPoses": numero totale di blendshape.
- "numFrames": numero totale di frame dell'animazione.
- "facsNames": array con i nomi di tutti i blendshape.
- "weightMat": matrice 2D con i pesi di ogni blendshape per ogni frame.

Di seguito un esempio di file JSON e della sua struttura: 4.1

```
{
    "exportFps": 60.0,
    "trackPath": "path/to/audio/file.wav",
    "numPoses": 59,
    "numFrames": 109,
    "facsNames": ["mouthRollUpperShape", "mouthCloseShape", ...],
    "weightMat": [[0.0, 0.0, 0.1846066564321518, ...], ...]
}
```

Figura 4.1. Esempio della struttura del file JSON

Importazione in Maya

Per importare l'animazione in Maya, viene utilizzato uno script Python. Questo script:

1. Legge il file JSON esportato.
2. Estrae le informazioni sui blendshape e i loro pesi.
3. Applica questi dati al modello in Maya, creando keyframes per ogni blendshape in ogni frame dell'animazione.

Il codice Python per l'importazione è strutturato come segue: 4.2

Questo approccio si rivela particolarmente vantaggioso per progetti che richiedono animazioni facciali dettagliate e personalizzabili, come film d'animazione, videogiochi AAA, o applicazioni di realtà virtuale e aumentata.

```
import maya.cmds as mc
import json

path = input("Insert path of json file:")

with open(path, "r") as f:
    facs_data = json.loads(f.read())
    facsNames = facs_data["facsNames"]
    numPoses = facs_data["numPoses"]
    numFrames = facs_data["numFrames"]
    weightMat = facs_data["weightMat"]

    bsNode = 'blendShape1'
    for fr in range(numFrames):
        for i in range(numPoses):
            mc.setKeyframe(bsNode+'.'+facsNames[i][:len(facsNames[i]) - 5], v=we
ightMat[fr][i], t=fr)
```

Figura 4.2. Codice Python per l'importazione dell'animazione su Autodesk Maya

4.2.3 Generazione di lipsync su un modello che non presenta dei blendshape

Questo approccio si distingue dal precedente principalmente nella fase di lavorazione all'interno di Audio2Face, dove si affronta la sfida di generare automaticamente i blendshape per modelli 3D che ne sono privi. Le fasi di preparazione in Maya, esportazione del file JSON e importazione in Maya rimangono sostanzialmente invariate rispetto al metodo descritto nella sezione precedente.

Di seguito andremo quindi ad analizzare il solo processo di generezione automatica dei blendshape in Audio2Face:

Come visto in precedenza nella fase che descrive l'Importazione in Audio2Face, utilizzando il tool Character Transfer di Audio2Face, si procede con l'allineamento del modello importato con il modello di riferimento interno del software.

Una volta completato il Character Transfer, si accede alla funzionalità Blendshape Generation. L'utente può scegliere tra due template predefiniti:

- NV 46 Pose: set standard di NVIDIA con 46 espressioni facciali.
- ARKit(1) 51 Pose: set compatibile con ARKit di Apple, comprendente 51 pose.

Il sistema analizza la topologia del modello importato e trasferisce i blendshape dal template scelto, adattandoli alla morfologia specifica del modello. Infine, è possibile esaminare i blendshape generati e, se necessario, modificarli o effettuare nuovamente il processo di trasferimento migliorando le corrispondenze tra i punti di controllo.

Questo approccio offre grande flessibilità, permettendo di generare rapidamente un set completo di blendshape anche per modelli che originariamente ne erano privi, e di integrarli facilmente in flussi di lavoro esistenti per l'animazione facciale.

Capitolo 5

Esperimenti

n questo capitolo verranno descritti gli esperimenti condotti per valutare l'efficacia della generazione di lipsync utilizzando Audio2Face. Gli esperimenti si sono focalizzati sui tre casi di studio analizzati in precedenza:

1. Live Streaming dei blendshape da Audio2Face ad Unreal Engine su un modello MetaHuman.

2. Generazione di lipsync su un modello che presenta dei blendshape preesistenti e import su Autodesk Maya.

3. Generazione di lipsync su un modello privo di blendshape e import su Autodesk Maya.

L'obiettivo è stato quello di analizzare e confrontare questi diversi approcci in termini di qualità del lipsync ottenuto, efficienza del workflow e flessibilità d'uso in diversi scenari applicativi.

5.1 Materiali e configurazioni

5.1.1 Software utilizzati

- Polycam: per l'acquisizione fotogrammetrica.
- Blender: per l'ottimizzazione delle mesh.
- Unreal Engine 5: per il rendering in tempo reale del MetaHuman.

- Autodesk Maya: per la preparazione dei modelli e l'importazione dell'animazione.
- NVIDIA Audio2Face: per la generazione del lipsync.

5.1.2 Configurazione delle macchine

- CPU: i3-9100f quad-core da 3.6Ghz
- GPU: Nvidia RTX 3060 12gb
- RAM: 16gb ddr4

5.2 Procedure sperimentali

In questa sezione verranno mostrate in dettaglio le procedure sperimentali adottate per ciascuno dei tre casi di studio analizzati. Ogni esperimento è stato progettato per valutare un aspetto specifico della generazione di lipsync utilizzando Audio2Face, con l'obiettivo di coprire una vasta gamma di scenari applicativi.

5.2.1 Esperimento 1: Live Streaming dei blendshape

L'esperimento si articola in due fasi principali: la creazione di un modello MetaHuman personalizzato e la configurazione del sistema di Live Streaming.

Creazione del modello MetaHuman

1. Acquisizione dati fotogrammetrici del soggetto utilizzando Polycam
    - Sono state acquisite 150 immagini del soggetto a 360°. Renderizzate con la modalità di elaborazione "Dettaglio Completo" che garantisce la realizzazione di una mesh ad alto dettaglio.
    - È stato poi effettuato l'export della mesh in formato .GLTF
2. Ottimizzazione della mesh risultante in Blender 5.1
    - In questa fase è stata effettuata una pulizia della mesh risultante attraverso le seguenti fasi:

- Applicazione del modificatore Shade Smooth
- Rimozione di vertici superflui
- Pulizia della geometria (Delete Loose, Degenerate Dissolve, Merge By Distance)

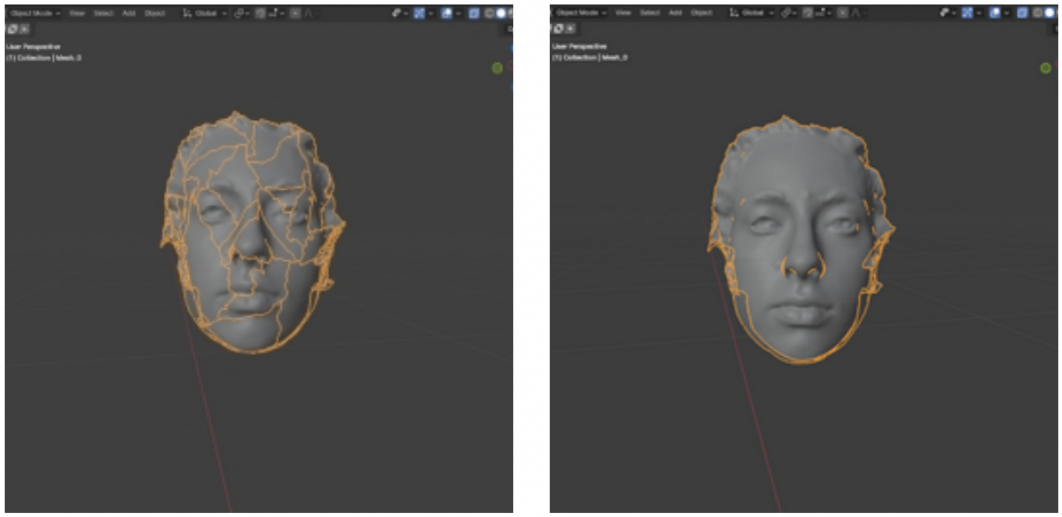


Figura 5.1. Mesh prima e dopo il clean-up

- Infine, l'export è stato effettuato con le seguenti specifiche 5.2

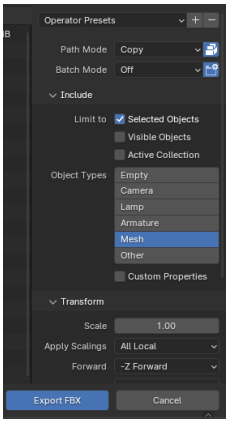


Figura 5.2. Impostazioni di export da Blender

3. Importazione in Unreal Engine e creazione della MetaHuman Identity 5.3

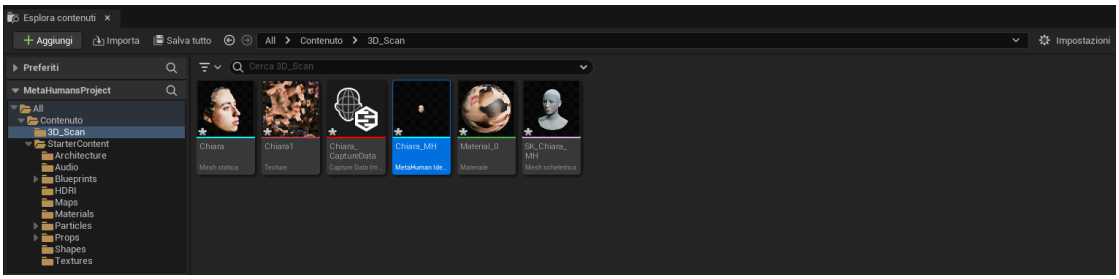


Figura 5.3. Creazione della MetaHuman Identity

- In questa fase sono stati mappati i landmark facciali della mesh sulla MetaHuman Identity 5.4

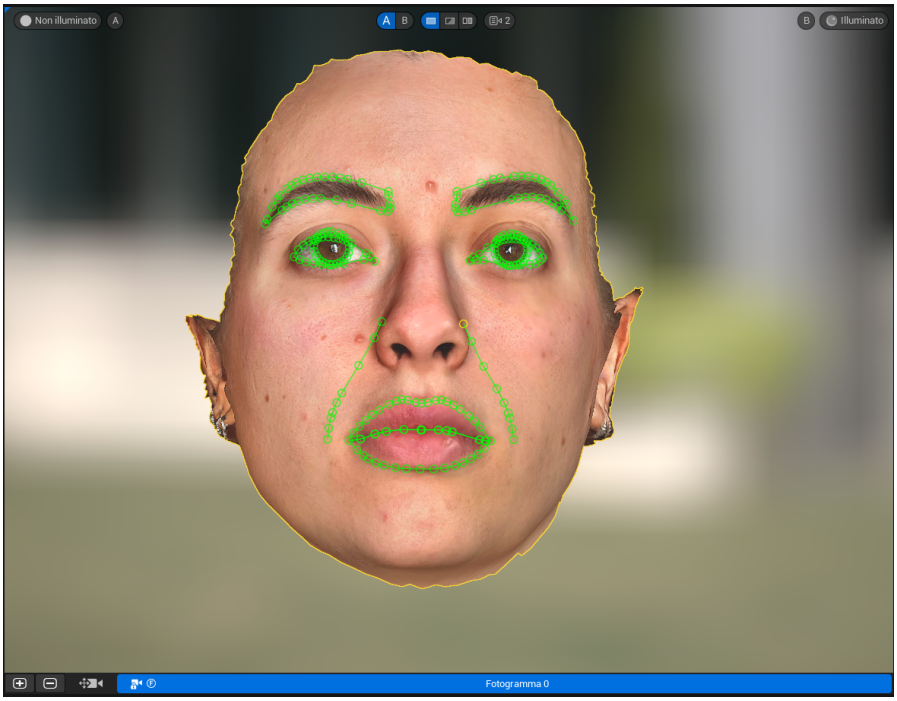


Figura 5.4. Mapping dei landmark facciali

- Viene effettuato il MetaHuman Identity Solve che permette di creare un MetaHuman basato su una scansione 3D o su una mesh di un

volto reale. Funziona analizzando la geometria e le caratteristiche del volto fornito e poi genera un MetaHuman che assomiglia il più possibile a quel volto.

- Infine si procede con il Mesh to MetaHuman che consente di convertire la mesh in un personaggio MetaHuman.

4. Rifinitura del modello nel MetaHuman Creator

   - In questa fase è stata effettuata una customizzazione del modello per renderlo il più somigliante possibile al soggetto scansionato.

     Di seguito è possibile vedere il modello prima e dopo la personalizzazione: 5.5 5.6

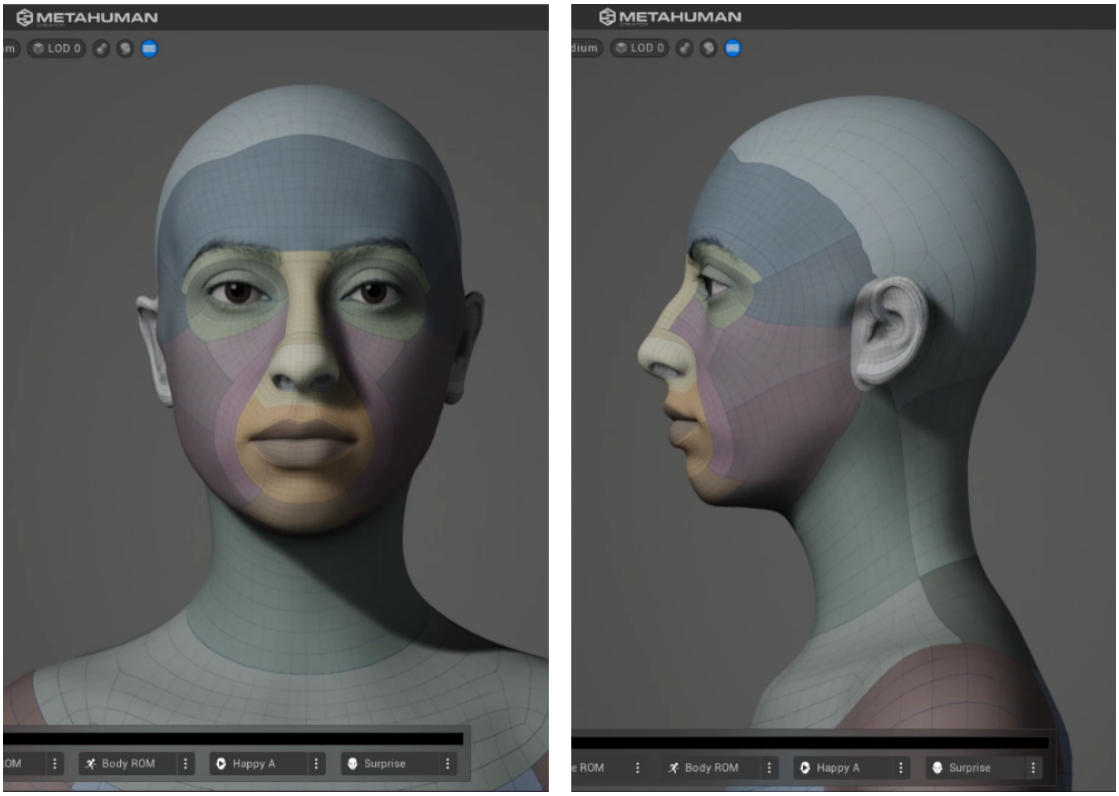


Figura 5.5. Modello MetaHuman prima della personalizzazione

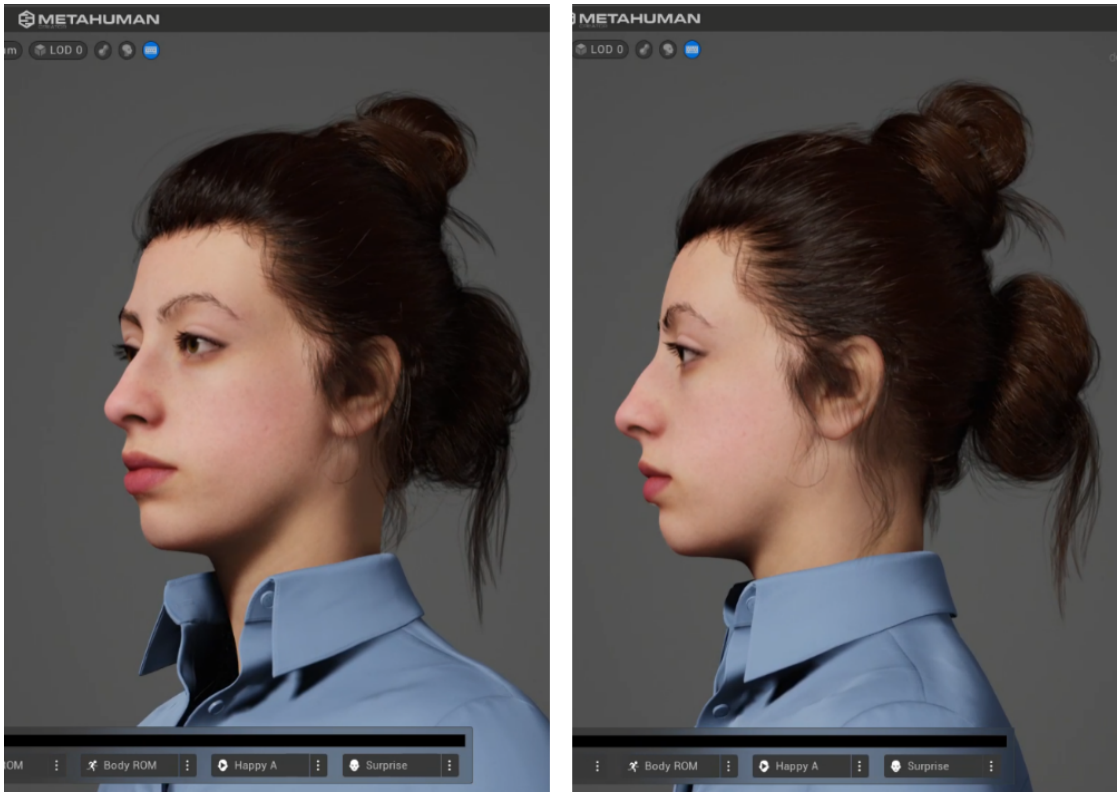


Figura 5.6. Modello MetaHuman dopo la personalizzazione

Configurazione del Live Streaming

1. Il processo comincia con l'installazione del plugin Audio2Face in Unreal Engine

2. Successivamente è stato configurato il LiveLink per la comunicazione tra Audio2Face e Unreal 5.7

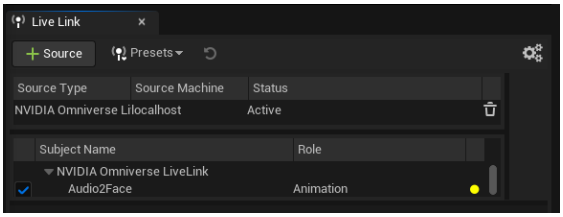


Figura 5.7. Configurazione del LiveLink da Audio2Face ad Unreal

3. Viene infine attivato lo streaming in tempo reale dei blendshape da Audio2Face 5.8

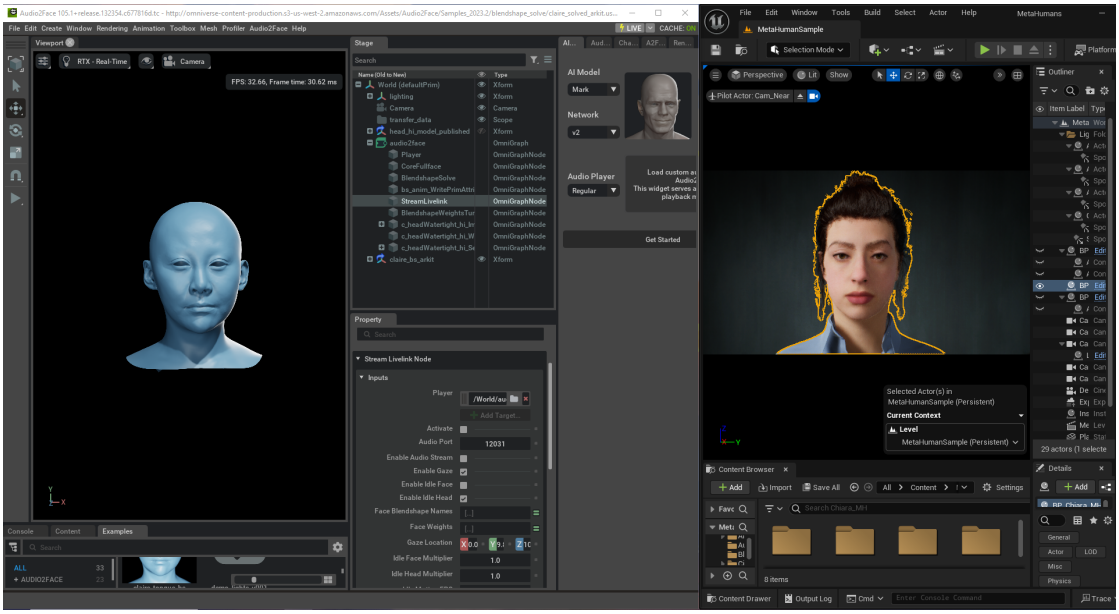


Figura 5.8. Live streaming dei blendshape da Audio2Face ad Unreal

5.2.2 Esperimento 2: Modello con blendshape

Questo esperimento si concentra sull'utilizzo di Audio2Face per generare lip-sync su un modello 3D che presenta già dei blendshape predefiniti. Questo approccio permette di sfruttare la precisione di blendshape creati manualmente, combinandola con l'efficienza e la rapidità della generazione automatica del lipsync.
La procedura seguita in questo esperimento mira a replicare un tipico workflow di produzione, dall'importazione del modello fino all'applicazione finale dell'animazione in Autodesk Maya.

1. Preparazione del modello 3D con blendshape predefiniti in Maya
2. Esportazione del modello in formato USD
    - Il modello viene esportato da Maya nel formato USD .Durante l'esportazione, ci si deve assicurare che tutte le informazioni relative ai blendshape siano correttamente incluse nel file USD. 5.9

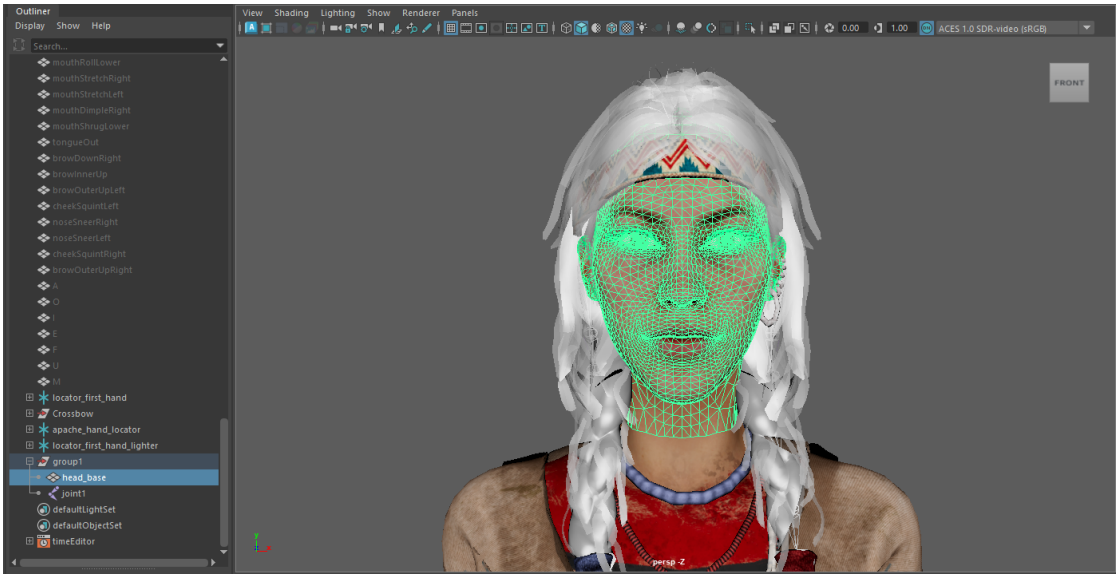


Figura 5.9. Preparazione del modello 3D in Maya

3. Importazione in Audio2Face e mappatura dei punti di controllo facciali 5.10

    - In questa fase si procede con il processo di mesh fitting, dove si allineano i punti di controllo del modello importato con quelli del modello di riferimento interno di Audio2Face. Questo passaggio è cruciale per garantire che Audio2Face possa manipolare correttamente i blendshape del modello importato. 5.11

4. Generazione dell'animazione lipsync in base all'input audio

    - In questa fase, si carica l'audio per il quale si vuole generare il lipsync. Audio2Face analizza l'audio utilizzando algoritmi di machine learning per interpretare i fonemi e le caratteristiche prosodiche. Basandosi su questa analisi, il sistema genera l'animazione facciale, manipolando i blendshape del modello importato per creare un lipsync realistico e sincronizzato con l'audio.

5. Tuning dei blendshape tramite Float Array Tuner

    - Dopo l'importazione dell'animazione, è stato effettuato un fine-tuning dei blendshape utilizzando il nodo Float Array Tuner 5.12.

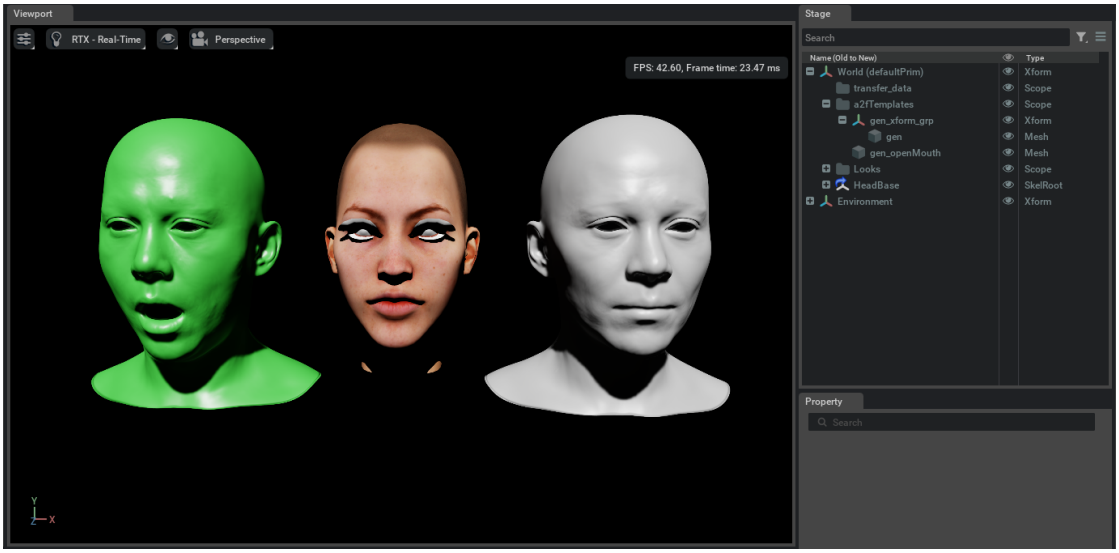


Figura 5.10. Import del modello su Audio2Face

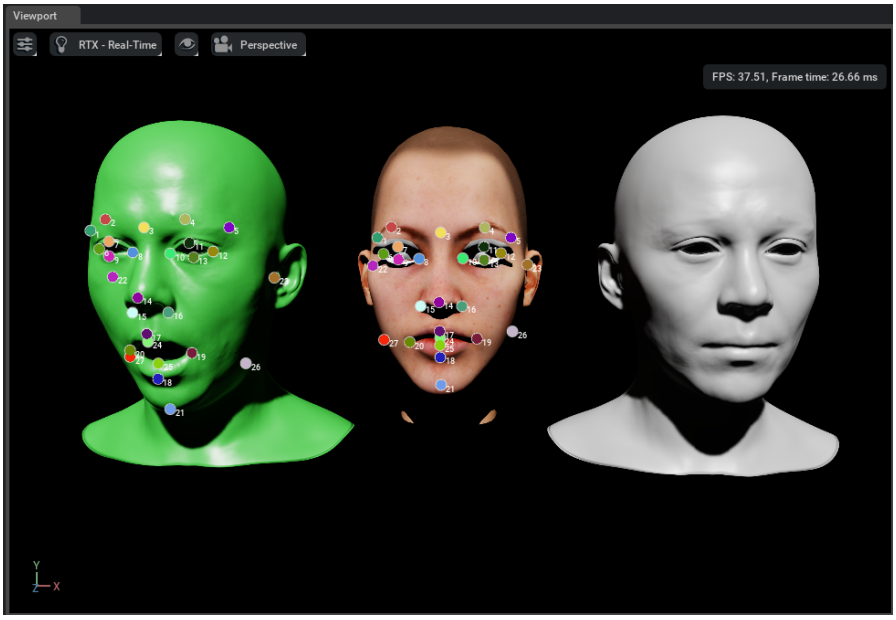


Figura 5.11. Mappatura dei punti di controllo facciali

Questo strumento permette di regolare con precisione i valori di guadagno (Gain) e offset per ciascun blendshape individualmente 5.13.

Il processo consente di calibrare l'intensità e la risposta di ogni espressione facciale, ottimizzando così il risultato finale del lipsync. I valori regolati vengono poi applicati attraverso il nodo Blendshape Solve, assicurando che l'animazione facciale sia perfettamente bilanciata e naturale.

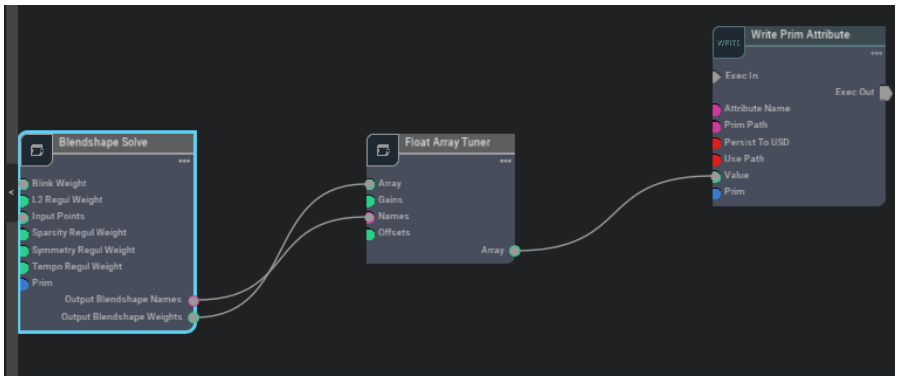


Figura 5.12. Nodo Float Array Tuner

6. Esportazione dei dati di animazione in formato JSON

   - Una volta generata l'animazione, i dati vengono esportati in un file JSON. Questo file contiene informazioni dettagliate su come ogni blendshape deve essere animato nel tempo, includendo i valori di peso per ogni frame dell'animazione.

7. Importazione e applicazione dell'animazione in Maya tramite script Python

   - Infine, si utilizza uno script Python personalizzato in Maya per importare i dati di animazione dal file JSON. Lo script legge i dati, li interpreta e li applica al modello originale in Maya, creando keyframes per ogni blendshape in ogni frame dell'animazione. Questo processo automatizza la creazione dell'animazione finale, permettendo eventuali regolazioni manuali da parte dell'animatore se necessario.

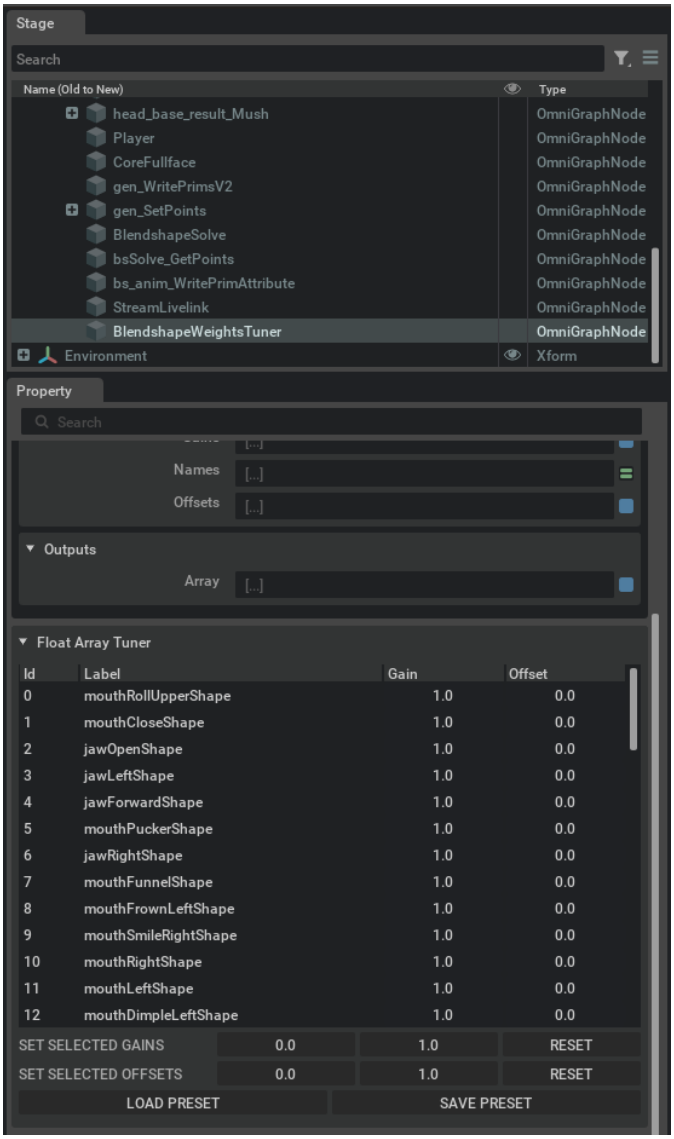


Figura 5.13. Blendshape Weights Tuner con Gain e Offset

5.2.3 Esperimento 3: Modello senza blendshape

Questo esperimento risulta essere uguale al precedente per le fasi 1-2, prendendo però in considerazione un modello privo di blendshape. L'esperimento procede nel seguente modo:

1. Generazione automatica dei blendshape

- Per avviare la generazione automatica dei blendshape, si accede alla funzionalità "Blendshape Generation" in Audio2Face. Qui si sceglie tra i template disponibili, NV 46 Pose o ARKit 51 Pose, in base alle esigenze del progetto.
- Una volta selezionato il template, si avvia il processo di generazione automatica. Audio2Face analizza attentamente la topologia del modello importato e crea un set completo di blendshape basandosi sul template scelto, adattandoli alla morfologia specifica del modello. 5.14

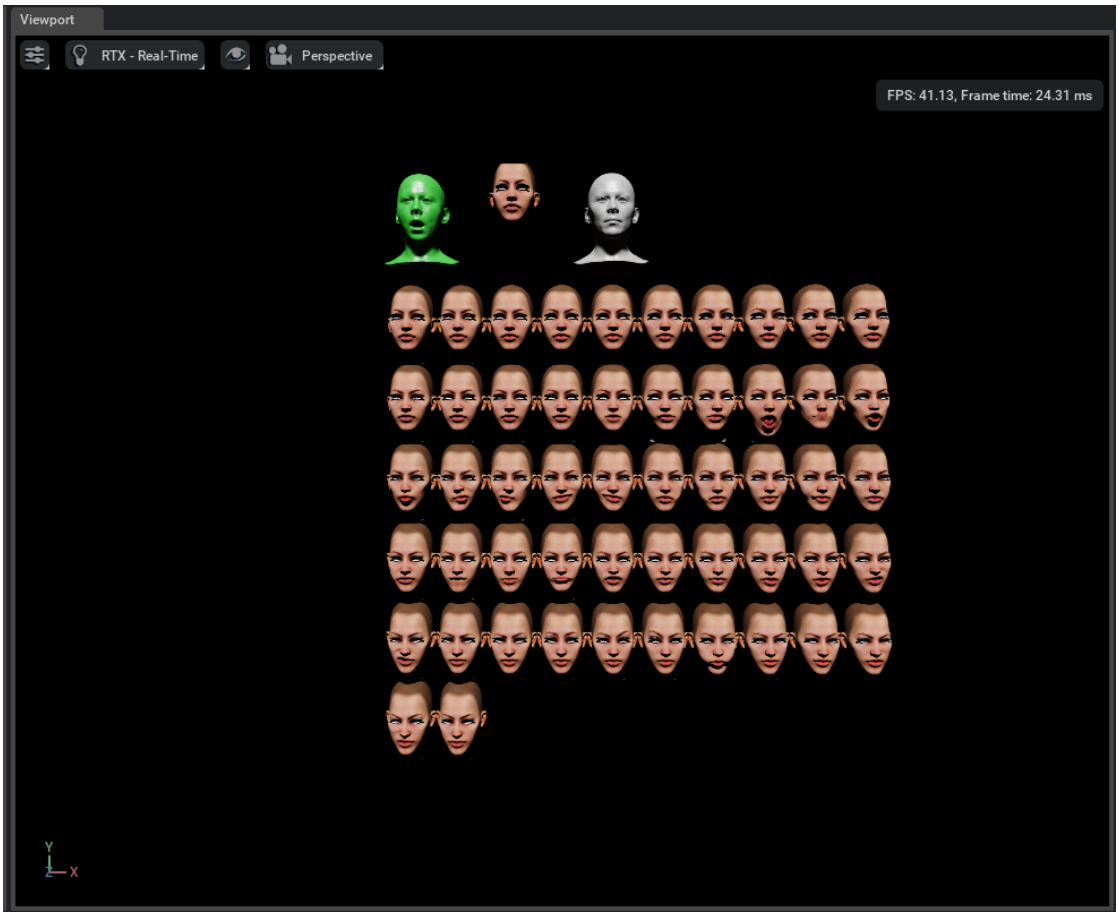


Figura 5.14. Set di blendshape generati da Audio2Face

2. Revisione e raffinamento dei blendshape generati

- Dopo la generazione, si procede con una revisione dettagliata di ogni blendshape creato. Si esamina attentamente ciascuno di essi per verificarne la qualità e l'accuratezza rispetto alle espressioni facciali desiderate.
- Nel caso in cui alcuni blendshape non soddisfino le aspettative, è possibile intervenire manualmente modificando i punti di corrispondenza o, se necessario, ripetere il processo di generazione per migliorare specifici blendshape che richiedono ulteriori perfezionamenti.

Per completare l'esperimento 3, si procede seguendo gli stessi passaggi dell'esperimento 2 dal punto 4 al punto 7. Questo approccio garantisce che, una volta generati i blendshape automaticamente, il processo di creazione del lipsync, tuning dell'animazione ed esportazione dei dati sia coerente tra i due esperimenti.

La principale differenza risiede nella fase iniziale di generazione automatica dei blendshape, che permette di ottenere risultati comparabili anche partendo da un modello privo di espressioni predefinite.

Capitolo 6

Risultati

In questo capitolo verranno riportati i risultati della valutazione effettuata sul modello di Text-to-Speech e su Audio2Face. I risultati sono stati ottenuti attraverso un form compilato da 11 partecipanti di varie fasce d'età e competenze in ambito linguistico e in animazione 3D.

Le valutazioni sono state fatte su:

- Comparazione Text-to-Speech: Bark e XTTS
- Valutazione lipsync e espressività emotiva di Audio2Face

6.1 Comparazione Text-to-Speech: Bark e XTTS

In questo studio, sono stati confrontati due modelli di sintesi vocale: Bark e XTTS. L'obiettivo era valutare la qualità della sintesi vocale prodotta da ciascun modello utilizzando frasi comunemente impiegate per testare le capacità dei sistemi TTS.

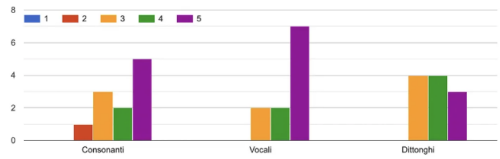
Le frasi valutate sono state:

- "The quick brown fox jumps over the lazy dog."
- "She sells seashells by the seashore."
- "How now, brown cow?"

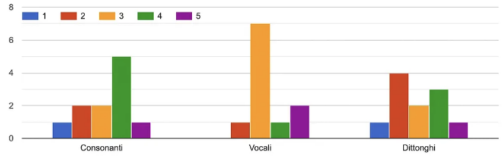
- "A proper copper coffee pot."
- "Betty Botter bought some butter."

Per ciascuna frase, sono state create due versioni sintetizzate, una per il modello Bark (**Versione A**) e una per il modello XTTS (**Versione B**). I partecipanti hanno valutato l'accuratezza delle consonanti, delle vocali e dei dittonghi per entrambe le versioni, utilizzando una scala da 1 (per niente accurato) a 5 (molto accurato).

Oltre all'accuratezza, è stata valutata anche la naturalezza percepita delle diverse versioni delle frasi. La naturalezza è stata definita come il grado in cui qualcosa appare o suona spontaneo, realistico e privo di artificiosità. I partecipanti hanno indicato quale versione (A o B) sembrava più naturale per ciascuna delle cinque frasi.



Versione A "The quick brown fox jumps over the lazy dog"



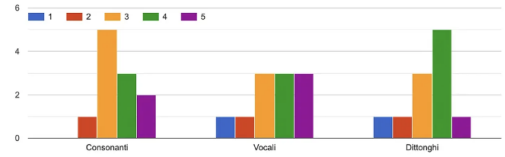
Versione B "The quick brown fox jumps over the lazy dog"

Figura 6.1. Valutazione consonanti, vocali e dittonghi della frase "The quick brown fox jumps over the lazy dog" Versione A e Versione B

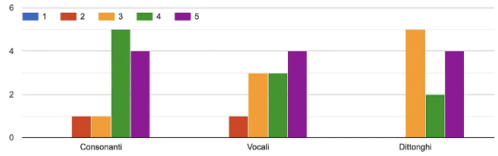
Per la frase "The quick brown fox jumps over the lazy dog", 10 partecipanti (90,9%) hanno ritenuto la Versione A più naturale rispetto alla Versione B. 1 partecipante (9,1%) ha ritenuto la Versione B più naturale rispetto alla Versione A. 6.1

Per la frase "She sells seashells by the seashore", 2 partecipanti (18,2%) hanno ritenuto la Versione A più naturale rispetto alla Versione B. 9 partecipanti (81,8%) hanno ritenuto la Versione B più naturale rispetto alla Versione A. 6.2

Per la frase "How now, brown cow?", 3 partecipanti (27,3%) hanno ritenuto la Versione A più naturale rispetto alla Versione B. 8 partecipanti (72,7%) hanno ritenuto la Versione B più naturale rispetto alla Versione A. 6.3

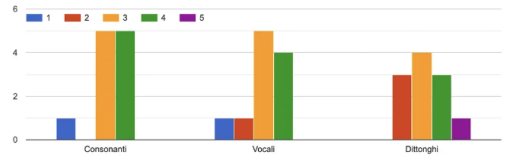


Versione A “She sells seashells by the seashore”

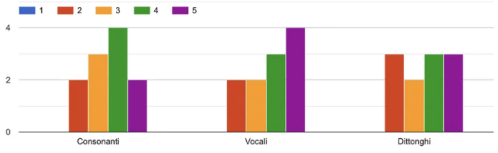


Versione B “She sells seashells by the seashore”

Figura 6.2. Valutazione consonanti, vocali e dittonghi della frase "She sells seashells by the seashore" Versione A e Versione B



Versione A “How now, brown cow?”



Versione B “How now, brown cow?”

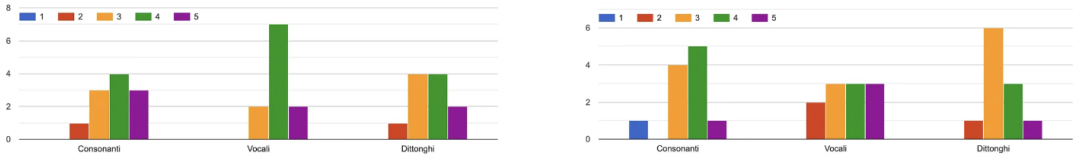
Figura 6.3. Valutazione consonanti, vocali e dittonghi della frase "How now, brown cow?" Versione A e Versione B

Per quanto riguarda la frase "A proper copper coffee pot", 7 partecipanti (63,6%) hanno ritenuto la Versione A più naturale rispetto alla Versione B. 4 partecipanti (36,4%) hanno ritenuto la Versione B più naturale rispetto alla Versione A. 6.4

Infine, per la frase "Betty Botter bought some butter", 5 partecipanti (45,5%) hanno ritenuto la Versione A più naturale rispetto alla Versione B. 6 partecipanti (54,5%) hanno ritenuto la Versione B più naturale rispetto alla Versione A. 6.5

Quindi, sebbene le preferenze siano state piuttosto bilanciate, il modello Bark (Versione A) sembra essere stato leggermente più apprezzato in termini di naturalezza dai partecipanti allo studio.

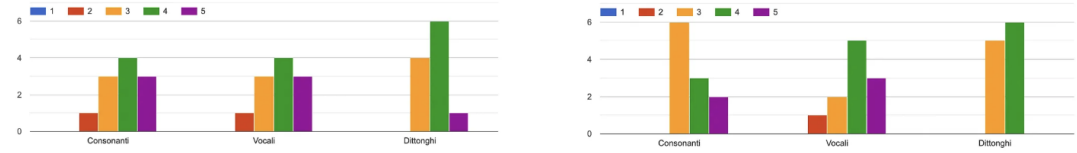
Tuttavia, va notato che si tratta di stime approssimative basate sui dati disponibili, e le differenze non sono così ampie da indicare una netta superiorità di un modello sull’altro nella produzione di sintesi vocale naturale. Entrambi i modelli sembrano avere punti di forza e debolezze specifiche a



Versione A "A proper copper coffee pot"

Versione B "A proper copper coffee pot"

Figura 6.4. Valutazione consonanti, vocali e dittonghi della frase "A proper copper coffee pot" Versione A e Versione B



Versione A "Betty Botter bought some butter"

Versione B "Betty Botter bought some butter"

Figura 6.5. Valutazione consonanti, vocali e dittonghi della frase "Betty Botter bought some butter" Versione A e Versione B

seconda dei contesti linguistici.

6.2 Valutazione lipsync e espressività emotiva di Audio2Face

Inoltre, è stata valutata la precisione del lipsync e il realismo delle espressioni facciali generate da un sistema di animazione audio2face. Sono stati presentati otto video in cui venivano rappresentate diverse espressioni facciali. I partecipanti hanno indicato quale emozione veniva rappresentata in ciascun video e hanno valutato quanto l'espressività emotiva fosse convincente, utilizzando una scala da 1 (per nulla convincente) a 5 (molto convincente).

L'emozione di Tristezza, presente nel Video 1, è stata riconosciuta dal 100% dei partecipanti.

A Fig.6.6 è possibile visionare la valutazione dell'espressività emotiva del Video 1.

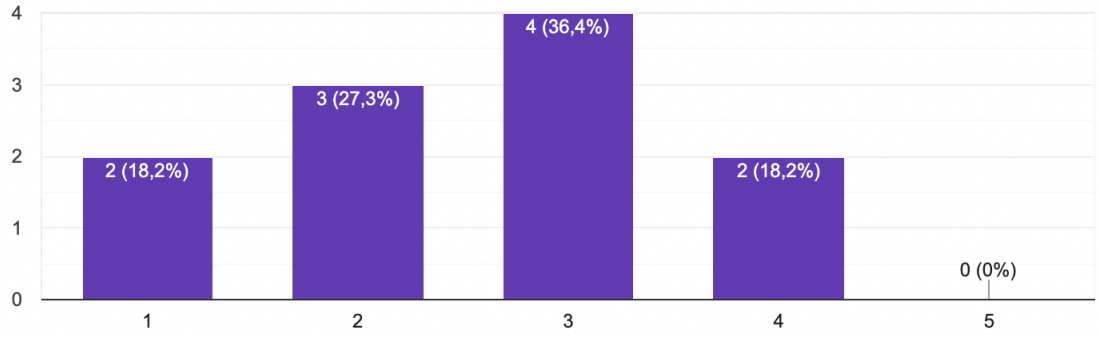


Figura 6.6. Valutazione dell'espressività emotiva del Video 1: Tristezza.

L'emozione di Rabbia, presente nel Video 2, è stata riconosciuta da 10 partecipanti (90,9%). 1 partecipante (9,1%) ha valutato l'emozione Non Riconoscibile.
A Fig.6.7 è possibile visionare la valutazione dell'espressività emotiva del Video 2.

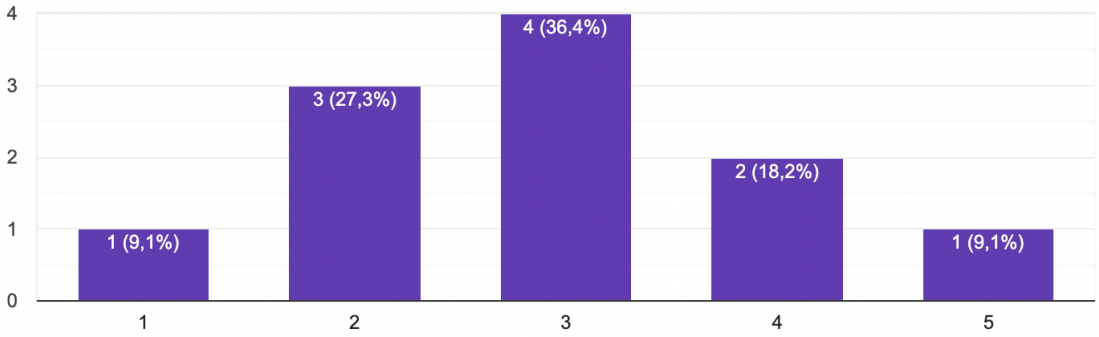


Figura 6.7. Valutazione dell'espressività emotiva del Video 2: Rabbia.

L'emozione di Sorpresa, presente nel Video 3, è stata riconosciuta da 7 partecipanti (63,6%). 3 partecipanti (27,3%) hanno valutato l'emozione come Rabbia. 1 partecipante (9,1%) ha valutato l'emozione Non Riconoscibile.

A Fig.6.8 è possibile visionare la valutazione dell'espressività emotiva del Video 3.

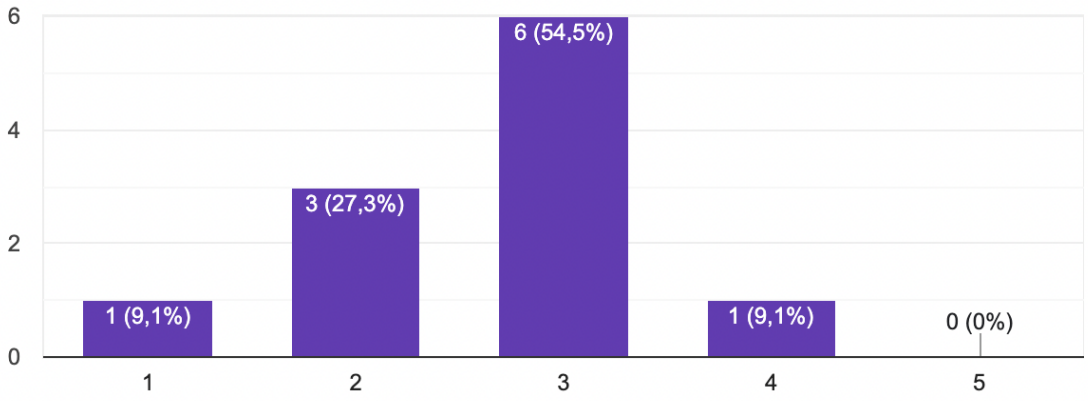


Figura 6.8. Valutazione dell'espressività emotiva del Video 3: Sorpresa.

L'emozione di Felicità, presente nel Video 4, è stata riconosciuta dal 7 partecipanti (63,6%). 3 partecipanti (27,3%) hanno valutato l'emozione Non Riconoscibile. 1 partecipante (9,1%) ha valutato l'emozione come Tristezza. A Fig.6.9 è possibile visionare la valutazione dell'espressività emotiva del Video 4.

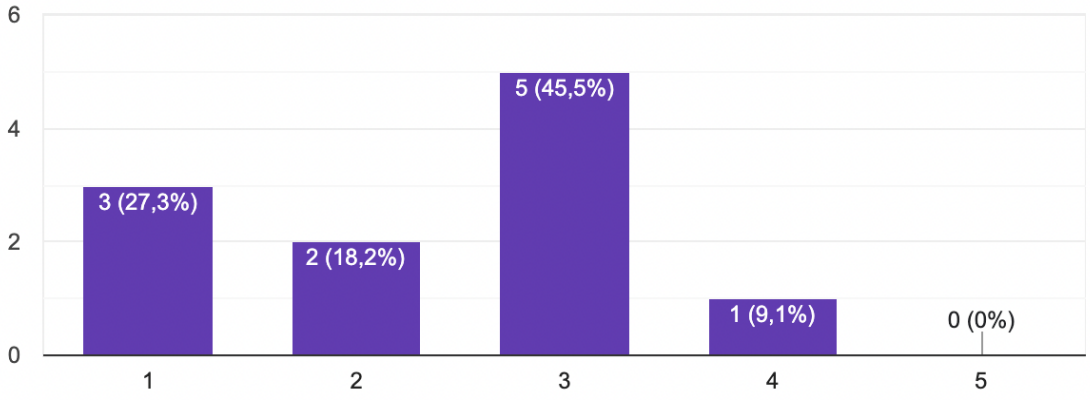


Figura 6.9. Valutazione dell'espressività emotiva del Video 4: Felicità.

L'emozione di Rabbia, presente nel Video 5, è stata riconosciuta da 8

partecipanti (72,7%). 3 partecipanti (27,3%) hanno valutato l'emozione Non Riconoscibile.
A Fig.6.10 è possibile visionare la valutazione dell'espressività emotiva del Video 5.

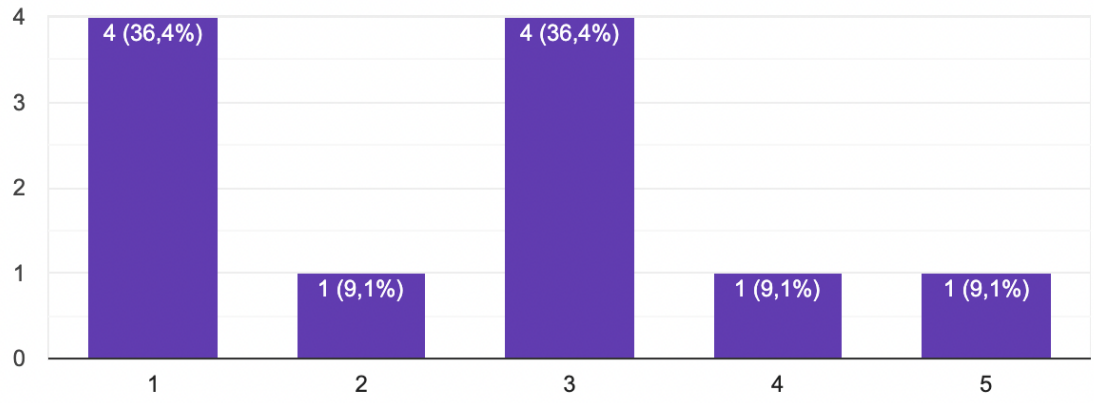


Figura 6.10. Valutazione dell'espressività emotiva del Video 5: Rabbia.

L'emozione di Felicità, presente nel Video 6, è stata riconosciuta da 8 partecipanti (72,7%). 3 partecipanti (27,3%) hanno valutato l'emozione Non Riconoscibile.
A Fig.6.11 è possibile visionare la valutazione dell'espressività emotiva del Video 6.

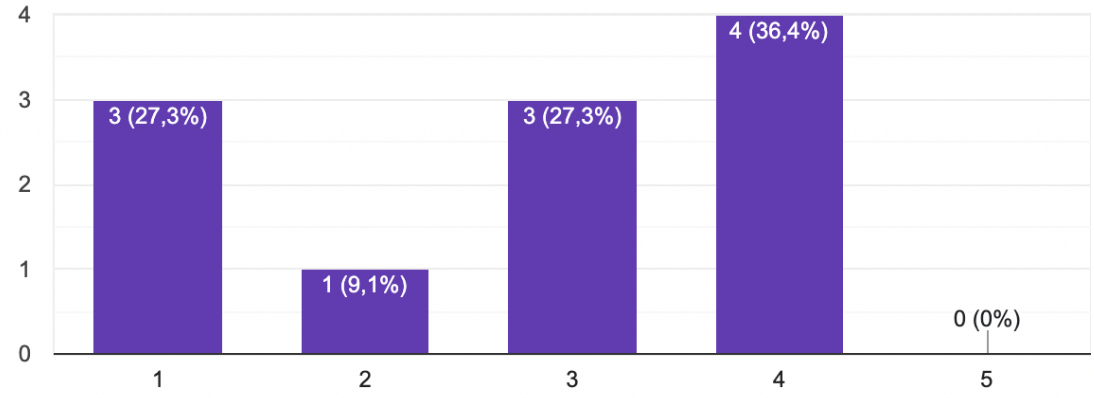


Figura 6.11. Valutazione dell'espressività emotiva del Video 6: Felicità.

L'emozione di Tristezza, presente nel Video 7, è stata riconosciuta da 9 partecipanti (81,8%). 2 partecipanti (18,2%) hanno valutato l'emozione Non Riconoscibile.
A Fig.6.12 è possibile visionare la valutazione dell'espressività emotiva del Video 7.

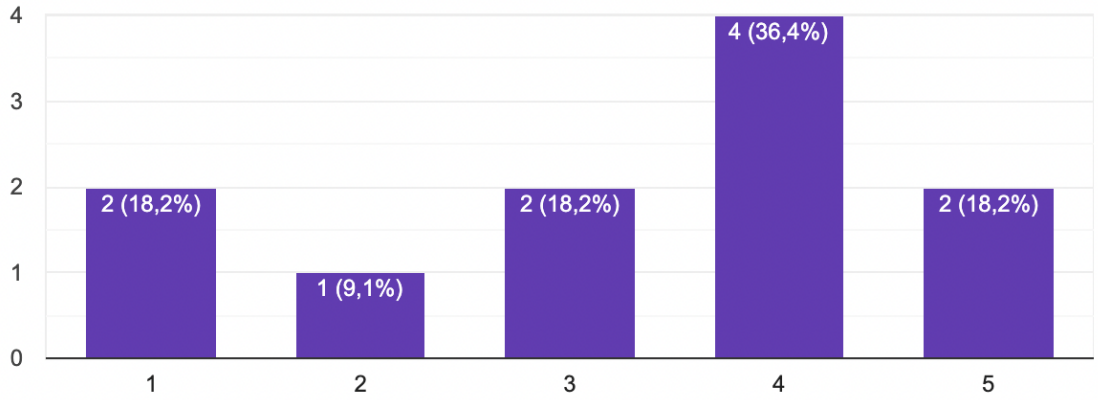


Figura 6.12. Valutazione dell'espressività emotiva del Video 7: Tristezza.

L'emozione di Sorpresa, presente nel Video 8, è stata riconosciuta da 9 partecipanti (81,8%). 1 partecipante (9,1%) ha valutato l'emozione come Rabbia. 1 partecipante (9,1%) ha valutato l'emozione Non Riconoscibile.
A Fig.6.13 è possibile visionare la valutazione dell'espressività emotiva del Video 8.

Sulla base dei dati presentati sulle valutazioni di convinzione dell'espressività emotiva e sul riconoscimento delle emozioni rappresentate nei video, possiamo formulare una valutazione complessiva delle prestazioni del sistema di animazione Audio2Face utilizzato in questo studio.

Nel complesso, i dati indicano che il sistema di animazione Audio2Face utilizzato in questo studio è in grado di produrre un'espressività emotiva moderatamente convincente, con prestazioni migliori per alcune emozioni rispetto ad altre. Tuttavia, vi è ancora spazio per miglioramenti al fine di aumentare il realismo e la riconoscibilità delle espressioni facciali generate.

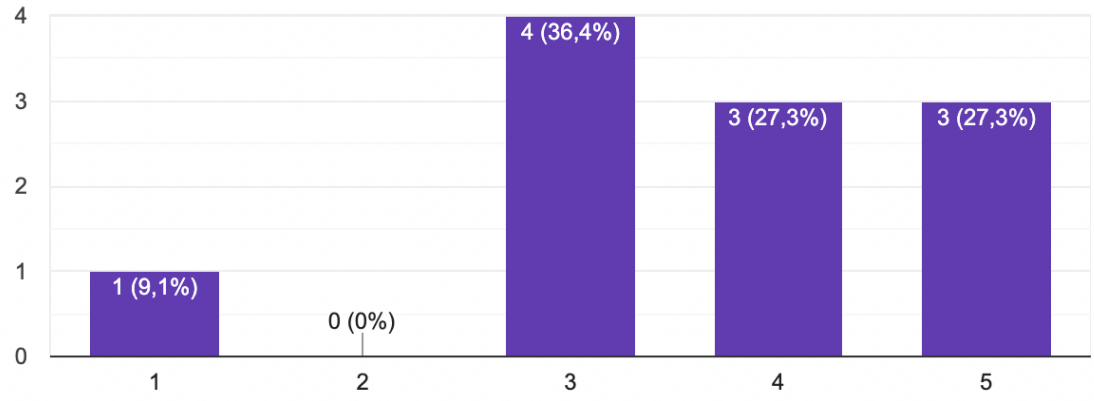


Figura 6.13. Valutazione dell'espressività emotiva del Video 8: Sorpresa.

Capitolo 7

Conclusioni e sviluppi futuri

7.1 Conclusioni

Questo studio ha esplorato l'applicazione di tecnologie avanzate di intelligenza artificiale nel campo dell'animazione facciale e della sintesi vocale, con particolare focus sull'utilizzo di NVIDIA Audio2Face per la generazione di lipsync e espressioni facciali realistiche.

I risultati ottenuti dagli esperimenti condotti mostrano che:

1. La pipeline integrata di Text-to-Speech e animazione facciale con Audio2Face sviluppata offre un workflow efficiente per la creazione di contenuti animati sincronizzati con l'audio. Il confronto tra i modelli Bark e Coqui ha evidenziato prestazioni comparabili, con una leggera preferenza per Bark in termini di naturalezza percepita.

2. Audio2Face si è dimostrato uno strumento potente e versatile per la generazione di lipsync, capace di gestire efficacemente sia modelli con blendshape predefiniti che modelli privi di essi. La possibilità di generare automaticamente i blendshape rappresenta un notevole vantaggio in termini di flessibilità e efficienza del workflow.

3. La valutazione dell'espressività emotiva generata da Audio2Face ha mostrato risultati promettenti, con un buon livello di riconoscibilità per la

maggior parte delle emozioni testate. Tuttavia, sono emersi anche margini di miglioramento, specialmente nella resa di alcune espressioni più complesse.

Nel complesso, lo studio ha dimostrato il potenziale delle tecnologie di IA nell'ottimizzare e migliorare i processi di animazione facciale, offrendo soluzioni che combinano efficienza, qualità e flessibilità.

7.2 Sviluppi futuri

Sulla base dei risultati ottenuti e delle limitazioni osservate, si possono delineare diverse direzioni per futuri sviluppi:

1. Miglioramento dell'espressività emotiva: affinare i parametri di Audio2Face per una resa ancora più realistica e sfumata delle espressioni facciali, specialmente per le emozioni risultate meno riconoscibili negli esperimenti.

2. Estensione del dataset di voci: Ampliare il dataset utilizzato per il Text-to-Speech, includendo una maggiore varietà di voci, accenti e stili di parlato per aumentare la versatilità del sistema.

3. Valutazione su larga scala: Condurre test più estesi coinvolgendo un campione più ampio e diversificato di utenti per valutare la percezione dell'animazione generata in diversi contesti culturali e applicativi.

Bibliografia

[1] Apple Inc. Arkit. URL https://developer.apple.com/augmented-reality/arkit/.

[2] Autodesk. Maya. URL https://www.autodesk.it/products/maya/overview?term=1-YEAR&tab=subscription.

[3] Blender Foundation. Blender. URL https://www.blender.org/.

[4] Hugging Face. Hugging face. URL https://huggingface.co/.

[5] NVIDIA. Audio2face, . URL https://www.nvidia.com/it-it/ai-data-science/audio2face/.

[6] NVIDIA. Audio-driven facial animation by joint end-to-end learning of pose and emotion, . URL https://research.nvidia.com/sites/default/files/publications/karras2017siggraph-paper_0.pdf.

[7] Pixar Animation Studios. Pixar. URL https://www.pixar.com/.

[8] Polycam Inc. Polycam. URL https://poly.cam/.

[9] Streamlit. Streamlit. URL https://streamlit.io/.

[10] Unreal Engine. Metahuman, . URL https://www.unrealengine.com/en-US/metahuman.

[11] Unreal Engine. Unreal engine, . URL https://www.unrealengine.com/en-US.